AD__________________

Award Number: DAMD17-00-1-0197

TITLE: Computer-Aided Diagnosis of Digital Mammograms

PRINCIPAL INVESTIGATOR: Yulei Jiang, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: June 2004

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY ***(Leave blank)***	***2. REPORT DATE*** June 2004	***3. REPORT TYPE AND DATES COVERED*** Annual Summary (1 Jun 2000 - 31 May 2004)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis of Digital Mammograms

5. FUNDING NUMBERS
DAMD17-00-1-0197

6. AUTHOR(S)
Yulei Jiang, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

E-Mail: y-jiang@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques to improve the detection and diagnosis of breast cancer. We have developed a computer technique that can classify breast calcifications in mammograms accurately, and this technique as a diagnostic aid has been shown to be able to improve radiologists' diagnostic accuracy. We have determined that Breast Imaging Report and Data System (BI-RADS) lesion descriptions provided by radiologists can be used as supplemental data to computer-extracted image features to improve the performance of computer classification of malignant and benign breast lesions. We have also found that our computer classification technique developed on screen-film mammograms, can achieve equally high performance on full-filed digital mammograms. This high performance is little affected by variability in the way in which radiologists indicate the general location of calcifications to the computer, which is designed as a means for the radiologist to query the computer aid. These results suggest that the computer technique has the potential to become a clinically useful and viable tool for diagnostic mammography.

14. SUBJECT TERMS
Computer-aided diagnosis (CAD), full-field digital mammography, BI-RADS, ROC analysis, artificial neural network (ANN)

15. NUMBER OF PAGES
61

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	***18. SECURITY CLASSIFICATION OF THIS PAGE*** Unclassified	***19. SECURITY CLASSIFICATION OF ABSTRACT*** Unclassified	***20. LIMITATION OF ABSTRACT*** Unlimited

TABLE OF CONTENTS

INTRODUCTION

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques to improve the detection and diagnosis of breast cancer. The hypothesis to be tested in the present project is that radiologists' ability to differentiate malignant from benign breast lesions can be improved by integrating radiologists' perceptual expertise in the interpretation of mammograms with the advantages of automated computer classification. This project has three SOW Tasks:

Task 1. To combine radiologist-extracted Breast Imaging Reporting and Data System (BI-RADS) features with image features extracted by a computer to classify malignant and benign clustered microcalcifications in mammograms.

Task 2. To optimally combine radiologists' diagnosis with the result of computer classification.

Task 3. To optimize computer classification for full-field digital mammograms.

BODY

1. Investigation related to BI-RADS

We reported last year that we have found radiologists-provided description of breast calcifications in mammograms using the BI-RADS lexicon tends to improve the performance of computer classification of calcifications as malignant or benign. However, variability among radiologists' use of the BI-RADS tends to diminish this gain in performance. Specifically, we have found little improvement in performance when a computer technique trained on one radiologist's BI-RADS data is tested on a different radiologist's BI-RADS data. This implies that the approach of adding radiologists-provided BI-RADS lesion descriptions to computer analysis of the lesion may be of limited practical value because of variability in radiologists' use of the BI-RADS lesion-description lexicon. We have stopped further work on this topic. However, we remain interested in the clinical use of BI-RADS. In particular, in interfacing our computer technique for classifying breast calcifications as malignant or benign (SOW Task 3) with radiologists, we present the clinical task in terms of BI-RADS assessment categories, as in typical clinical practice. In addition, we are interested in the effect (and perhaps the validity) of using BI-RADS final assessments as categorical data for ROC analysis, as is a common practice in research. No specific results to report at this time.

2. An analytical comparison of four methods for combining multiple sources of diagnostic information

This particular work is within the scope of SOW Task 2. Last year we reported a work on an analytical comparison—based on receiver operating characteristic (ROC) analysis [1]—of four methods for combing multiple diagnostic assessments of the same patient. That work was motivated by *ad hoc* use of various simple methods in CAD research, such as taking the simple average [2], or taking the result of the one image that is most indicative of a disease outcome (e.g., malignancy) [3]. The objective of that work was to understand the merit of these methods from a theoretical perspective. In the work reported last year we made two assumptions. One assumption was that diagnostic information derived from multiple images of the same patient can be described by the same binormal ROC curve [4]; the second assumptions was that diagnostic information derived from multiple images of the same patient is uncorrelated. We have continued this work by eliminating the second assumption to make the work more general: multiple images of the same patients such as mediolateral-oblique (MLO) and craniocaudal (CC) view mammograms are generally almost always correlated. We have found that our previous results hold in a special situation when correlation strength is zero. Those previous results were that the method of the simple average of data from multiple images of the same patient always produces an improved ROC curve over data from a single image. However, the method of taking data from the one image that is most indicative of malignancy and even the method of taking data from the one image that is least indicative of malignancy can also improve the ROC curve and, under certain conditions, outperform the method of the simple average to become the preferred method. Results taking correlation into account are similar. An important new finding is that as correlation strength increases, it becomes less often that the method of simple average produces the best ROC curve and it becomes more often that the method of taking data from the one image that is most, or least, indicative of malignancy produces the best ROC curve. These new findings are more general than previously reported and, therefore, more practically instructive. We will continue this work by removing the remaining assumption to make the work even more general. The earlier work was presented at the

RSNA meeting in 2003 [Liu, 2005 #37]. A peer-reviewed publication of that work appears in *medical Physics* [6]. The new work on combining correlated diagnostic decision variables will be presented at the RSNA meeting in 2004 [7] and the SPIE Medical Imaging Conference in 2005 [Liu, 2005 #47]. A manuscript describing this work is in final preparation and will be submitted to *medical physics.*

3. Investigation of a quadratic method for combining multiple sources of diagnostic information

This particular work is within the scope of SOW Task 2. In 2002 we reported a work developing an "optimal" method for combining quantitative diagnostic assessments made by a radiologist and by a computer, based on a bivariate binormal model that was originally developed for ROC analysis [9]. This method takes into account the individual accuracy of the radiologist and the computer, as well as the correlation between their diagnostic assessments. This method is optimal if the bivariate binormal model is appropriate for describing the underlying data, which we expect to be generally true, and if there are enough data to estimate model parameters accurately. This method is referred to as quadratic averaging (or quadratic for short) because it involves quadratic terms (i.e., to the power of 2) in averaging decision variables. Previously we claimed this method to be optimal based on theoretical grounds, under the conditions stated above [10]. We have now shown experimentally that this claim is indeed true. To demonstrate "optimality", it is necessary to calculate the ideal observer performance. However, the ideal observer performance is difficult to calculate analytically because it involves elliptical integration. We have now circumvented this difficulty by calculating the ideal observer performance numerically from a large sample of data in simulation studies. We have shown that as the amount of data increases, results of the quadratic method asymptotically approach performance of the ideal observer. A manuscript describing this work is in its final stage of preparation and will be submitted to *academic Radiology.*

4. Computer-aided diagnosis of malignant and benign calcifications in full-field digital mammograms

This particular work is within the scope of SOW Task 3. Last year we reported an evaluation on full-field digital mammograms of a computer technique that classifies calcifications in mammograms as malignant or benign that we developed previously on digitized screen-film mammograms. That study was significant because it was an independent evaluation in that: (1) the computer technique was developed on digitized screen-film images and evaluated on full-field digital images, (2) the computer technique was developed on older cases and evaluated on newer, completely different, cases, and (3) the computer technique was developed based on manual identification of individual calcifications and evaluated based on automatic detection of individual calcifications by the computer with radiologists' input limited to the general location of a group of calcifications. We found that the computer technique achieved virtually the same performance on full-field digital mammograms as on digitized screen-film mammograms that we reported previously in the literature. Further, the computer technique achieved highly consistent performance despite variability in radiologists' performance and in their input to the computer. Those results were obtained on 49 cases of calcification lesions that were deemed clinically suspicious for malignancy and biopsied. We have now performed a similar analysis on cases of calcifications that were clinically not biopsied and have found similar results. The significance of the present study is that for our computer technique to help radiologists reduce the number of biopsies performed on benign calcifications, it is necessary for the computer technique to advice radiologists not to biopsy some truly benign calcifications that radiologists may consider suspicious for malignancy, and also consistently advice radiologists not to biopsy truly benign calcifications that radiologists consider safe to follow. To further continue this research, we plan to use the computer technique to analyze consecutive mammograms of diagnostic studies (as opposed to screening mammograms) from one year at the University of Chicago. Results from that study should measure conclusively the performance of this computer technique. We have developed required logistic capabilities to access consecutive mammography cases, radiology reports, and pathology reports, and are well underway to carry out that study, perhaps in the next year. (It was a substantial task to develop the logistics for this study given the need to fulfill IRB and HIPPA requirements and the need to handle large number of patient cases.) This work was presented at the ARRS meeting in 2004 [11] and the International Digital Mammography Workshop in 2004 [12]. A manuscript for a peer-reviewed publication is under preparation.

5. A new method for training artificial neural networks

This particular work is beyond but related to SOW Task 3. We have continued investigating artificial neural networks. While not specifically stated in the SOW, artificial neural network is a key component of our computer technique for classifying breast calcifications as malignant or benign, and therefore, is related to SOW Task 3. We have developed a new method of training artificial neural networks that improves generalizability of artificial neural networks from training (i.e., to learn about cases at large rather than to learn about the specific cases in a training dataset) and also improves the performance of artificial neural networks. This method is known as training with "jitter". It introduces random noise into the training data, effectively making data points "jitter" in the abstract space of the image feature data, when training artificial neural networks. The added noise discourages the neural network from learning specifically the cases in the training dataset, and effectively enlarges the size of the training dataset substantially, both effects of which help the neural network to become more generalizable to cases at large. We have found that artificial neural networks trained with this method can achieve an improved A_z value (area under the ROC curve) of 0.88 compared to an A_z value of 0.80 from an artificial neural network trained with the conventional method in a previous study [13]. This work was presented at the RSNA meeting in 2003 [14] and at the AAPM meeting and the CARS meeting in 2004 [15]. A proceeding paper describing part of this work is attached.

KEY RESEARCH ACCOMPLISHMENTS

- Determined that the combination of BI-RADS lesion descriptors provided by radiologists and image features extracted by a computer can improve the performance of computer classification of malignant and benign breast lesions in mammograms, but reader variability in providing the BI-RADS lesion descriptors can diminish that improvement.
- Demonstrated that the method of choice for simple un-weighted linear combinations of diagnostic information derived from multiple sources such as multiple images of the same patient is not always a single method but will change from one method to another depending on the ROC curve parameters of the diagnostic information derived from each single source, and depending on the correlation strength in the decision variables of diagnostic information derived from each single source.
- Demonstrated empirically that a quadratic averaging method we described previously for combining correlated diagnostic assessments, such as those made by a radiologist and made by a computer-aided

diagnosis technique on the same patient, produces results asymptotically approach the performance of the ideal observer, in support of a theoretical claim of the same conclusion.

- Demonstrated that our computer-aided diagnosis technique that classifies calcifications in mammograms as malignant or benign developed previously on digitized screen-film mammograms and required manual identification of individual calcifications can achieve virtually the same highly accurate and highly consistent performance on full-field digital mammograms.
- Evaluated the computer-aided diagnosis technique on both lesions biopsied because of suspicion of malignancy and lesions not biopsied because radiologists considered them safe to follow.
- Developed necessary logistics and infrastructure to conduct a large evaluation of the computer-aided diagnosis technique on consecutive diagnostic mammography studies from one year.
- Developed a novel technique for training artificial neural networks by using "jitter" that can improve generalizability and performance of the artificial neural networks.

REPORTABLE OUTCOMES

Manuscripts

1. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Newstead GM. Use of BI-RADS lesion descriptors in computer-aided diagnosis of malignant and benign breast lesions. *Proc SPIE* 5372 199-202, 2004.
2. Liu B, Metz CE, Jiang Y. An ROC comparison of four methods of combining information from multiple images of the same patient. *Medical Physics* 31:2552-2563, 2004.
3. Paquerault S, Yarusso LM, Papaioannou J, Jiang Y, Nishikawa RM. Radial gradient-based segmentation of mammographic microcalcifications: observer evaluation and effect on CAD performance. *Medical Physics* 31:2648-2657, 2004.
4. Jiang Y, Rana RS, Schmidt RA, Nishikawa RM, Liu B, Sennett CA, Chambliss J, Abe H. Computer classification of malignant and benign calcifications in full-field digital mammograms. In Pisano E, Ed., *Digital Mammography 2004*, (in press).

5. Nishikawa RM, Jiang Y, Reiser I. What is the required pixel size for digital mammography? In Pisano E, Ed., *Digital Mammography 2004*, (in press).

6. Zur RM, Jiang Y, Metz CE. Comparison of two methods of adding jitter to artificial neural network training. In Lemke HU, Vannier MW, Inamura K, Farman AG, Doi K, Reiber JHC Eds., *CARS 2004 Computer Assisted Radiology and Surgery*, Amsterdam: Elsevier, 886-889, 2004.

Abstracts

7. Liu B, Metz CE, Jiang Y. ROC comparison of three methods of analyzing information derived from multiple images of the same patient with application to computer-aided diagnosis (CAD). In: *Radiological Society of North America scientific assembly and annual meeting program.* Oak Brook, IL: Radiological Society of North America, 425-426, 2003.

8. Rana R, Jiang Y, Schmidt RA, Liu B, Sennett C, Chambliss J, Abe H, Lunning N. Independent evaluation of computer classification of malignant and benign calcifications in full-filed digital mammograms. *American Journal of Roentgenology* 182 (supplement):30, 2004.

9. Liu B, Jiang Y, Rana R. Effect of radiologists' variability on computer performance in classifying malignant and benign microcalcifications in mammograms. *Medical Physics* 31:1795, 2004.

10. Paquerault S, Yarusso LM, Nishikawa RM, Papaioannou J, Edwards AV, Jiang Y. Observer evaluation and CAD performance of a radial gradient-based segmentation method for mammographic microcalcifications. In: *Radiological Society of North America scientific assembly and annual meeting program.* Oak Brook, IL: Radiological Society of North America, 389, 2003.

11. Zur RM, Jiang Y. Avoiding overfitting and increasing generalizability of artificial neural networks in CAD by training with jitter. In: *Radiological Society of North America scientific assembly and annual meeting program.* Oak Brook, IL: Radiological Society of North America, 390, 2003.

12. Zur R, Jiang Y. Variability in the outputs of Bayesian artificial neural networks. *Medical Physics* 31:1795, 2004.

Non-Abstracted Presentations

13. Jiang Y. Digital mammography and computer-aided diagnosis. Invited presentation at the *Second Seoul International Symposium for Computer Aided Diagnosis*, Seoul, Korea, 2004.

14. Jiang Y. Computer-aided diagnosis of breast calcifications. Invited presentation at the *Second Seoul International Symposium for Computer Aided Diagnosis*, Seoul, Korea, 2004.

CONCLUSIONS

We have made significant progress toward completing all three SOW Tasks. All three Tasks now may be considered as complete. However, research addressed by the three SOW Tasks continues, particularly for Tasks 2 and 3. The continuing research is beyond the scope of the SOW of this grant and we believe it will lead to new insights and advances in the research questions it addresses. We plan to further pursue this line of research. We gratefully acknowledge the support of the US Army Breast Cancer Research Program, particularly in helping the PI establishing a career in breast cancer research.

REFERENCES

1. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 21:720-733, 1986.

2. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 212:817-827, 1999.

3. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ, Doi K. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 198:671-678, 1996.

4. Dorfman DD, Alf E, Jr. Maximum likelihood estimation of parameters of signal detection theory--a direct solution. Psychometrika 33:117-124, 1968.

5. Liu B, Metz CE, Jiang Y. ROC comparison of three methods of analyzing information derived from multiple images of the same patient with application to computer-aided diagnosis (CAD) (abstract). In: Radiological Society of North America scientific assembly and annual meeting program eds.). Oak Brook, IL: Radiological Society of North America, pp. 425-426, 2003.

6. Liu B, Metz CE, Jiang Y. An ROC comparison of four methods of combining information from multiple images of the same patient. Med Phys 31:2552-2563, 2004.

7. Liu B, Metz CE, Jiang Y. Effect of correlation on combining diagnostic information from two images of the same patient (abstract). In: Radiological Society of North America scientific assembly and annual meeting program eds.). Oak Brook, IL: Radiological Society of North America, pp. 2004.

8. Liu B, Metz CE, Jiang Y. A theoretical investigation of methods for combining multiple diagnostic assessments. Proc SPIE 2005.

9. Metz CE, Wang P-L, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Information Processing in Medical Imaging (Deconinck F, eds.). Nijhoff: The Hague, pp. 432-445, 1984.

10. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. *Proc. SPIE* 4324:177-183, 2001.

11. Rana R, Jiang Y, Schmidt RA, Liu B, Sennett C, Chambliss J, Abe H, Lunning N. Independent evaluation of computer classification of malignant and benign calcifications in full-filed digital mammograms. American Journal of Roentgenology 182 (supplement):30, 2004.

12. Jiang Y, Rana RS, Schmidt RA, Nishikawa RM, Liu B, Sennett CA, Chambliss JJ, Abe H. Computer classification of malignant and benign calcifications in full-field digital mammograms. In: Digital Mammography 2004 (Pasano E, eds.). (in press), 2004.

13. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 6:22-33, 1999.

14. Zur RM, Jiang Y. Avoiding overfitting and increasing generalizability of artificial neural networks in CAD by training with jitter (abstract). In: Radiological Society of North America scientific

assembly and annual meeting program eds.). Oak Brook, IL: Radiological Society of North America, pp. 390, 2003.

15. Zur R, Jiang Y. Variability in the outputs of Bayesian artificial neural networks (abstract). Medical Physics 31:1795, 2004.

LIST OF ATTACHED REPRINTS

1. Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ, Newstead GM. Use of BI-RADS lesion descriptors in computer-aided diagnosis of malignant and benign breast lesions. Proc SPIE 5372 199-202, 2004.
2. Liu B, Metz CE, Jiang Y. An ROC comparison of four methods of combining information from multiple images of the same patient. *Medical Physics* 31:2552-2563, 2004.
3. Paquerault S, Yarusso LM, Papaioannou J, Jiang Y, Nishikawa RM. Radial gradient-based segmentation of mammographic microcalcifications: observer evaluation and effect on CAD performance. *Medical Physics* 31:2648-2657, 2004.
4. Jiang Y, Rana RS, Schmidt RA, Nishikawa RM, Liu B, Sennett CA, Chambliss JJ, Abe H. Computer classification of malignant and benign calcifications in full-field digital mammograms. In Pisano E, Ed., *Digital Mammography 2004*, (in press).
5. Nishikawa RM, Jiang Y, Reiser I. What is the required pixel size for digital mammography? In Pisano E, Ed., *Digital Mammography 2004*, (in press).
6. Zur RM, Jiang Y, Metz CE. Comparison of two methods of adding jitter to artificial neural network training. In Lemke HU, Vannier MW, Inamura K, Farman AG, Doi K, Reiber JHC Eds., *CARS 2004 Computer Assisted Radiology and Surgery*, Amsterdam: Elsevier, 886-889, 2004.

Use of BI-RADS lesion descriptors in computer-aided diagnosis of malignant and benign breast lesions

Yulei Jiang, Robert A. Schmidt, Robert M. Nishikawa,
Carl J. D'Orsi*, Carl J. Vybomy, Gillian M. Newstead

Department of Radiology, The University of Chicago, Chicago, IL 60637

*Department of Radiology, Emory University, Atlanta, GA 30322

ABSTRACT

The purpose of this study was to determine whether combining an automated computer technique that classifies calcifications in mammograms as malignant or benign with radiologist-provided BI-RADS lesion description improves classification performance. Three expert mammography radiologists who were MQSA certified and familiar with BI-RADS retrospectively interpreted 125 cases of mammograms containing calcifications and provided BI-RADS lesion descriptions. A computer technique was applied to the mammograms to extract eight image features that describe the size, shape, and uniformity of individual as well as groups of calcifications. We compared the performance of artificial neural networks that estimated the likelihood of malignancy based on input from either the computer-extracted image features alone, the BI-RADS lesion descriptors alone, or the combination of both. The leave-one-out method was used. Combining the BI-RADS lesion description provided by a single radiologist and computer-extracted image features resulted in improved performance. However, using two radiologists' BI-RADS lesion descriptions such that one radiologist's data was used to train and another radiologist's data was used to test the neural network diminished this improvement in performance. These results suggest that variability in radiologists' BI-RADS lesion description is large enough to offset a potential gain in performance from combining it with an automated computer technique.

Keywords: BI-RADS, computer-aided diagnosis, classification, breast calcifications, reader variability

1. INTRODUCTION

Computer-aided diagnosis (CAD) is being developed to help radiologists improve their diagnostic performance in the interpretation of screening and diagnostic mammograms. Previous research demonstrates that computer-aided diagnosis holds the potential to help radiologists reduce the number of biopsies of benign lesions while maintaining or even increasing the correct diagnosis of malignant lesions [1-4]. Two different approaches have been taken in developing CAD techniques. The first is to rely on subjective description of breast lesions provided by radiologists in conjunction with a computer classifier to classify breast lesions [5, 6]. This approach is believed to help the radiologist by inducing the radiologist to observe the lesion and rate the lesion appearance in a systematic manor and by the use of a computer classifier that might be more apt than humans in analyzing more than a couple of lesion descriptors. The lesion description lexicon of the Breast Imaging Reporting and Data System (BI-RADS) has been used as a basis for this approach [7]. The second approach is to rely on the computer to extract image features that describe breast lesions in conjunction with a computer classifier to classify breast lesions [2, 4, 8]. This approach has the advantage that both image-feature extraction and classification decision-making are done in an objective way. Often, radiologists' subjective experience is used to guide the development of computer extraction of image features [8].

Previously, we have shown that by combining these two approaches, that is to use a computer classifier to analyze both BI-RADS lesion descriptors and computer-extracted image features, computer classification performance of malignant and benign breast lesions can be improved [9]. However, we cautioned that variability in radiologists' lesion

Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment,
edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.537183

descriptions would counteract this apparent benefit from combining the two sources of information on lesion appearance. Studies have measured variability among radiologists in describing breast lesions in terms of the BI-RADS lexicon, and have found generally only moderate agreement [10, 11]. The purpose of this work was to investigate whether it is beneficial to combine computer-extracted image features and BI-RADS lesion descriptors provided by radiologists for computer classification of calcifications as malignant or benign.

2. MATERIALS AND METHODS

2.1. Reader study

Three expert radiologists specializing in mammography read 125 cases of mammograms containing calcifications. These radiologists were familiar with the BI-RADS lexicon of lesion descriptors. In 41 cases the calcifications were associated with cancers. These were confirmed by biopsy. Benign cases were confirmed either by biopsy or by follow-up mammogram study. The radiologists read original mammograms in standard and magnification views of both breasts. The calcifications in question were indicated on all films. The radiologists read the cases in random order with no additional information, and with no limit on reading time.

For each case, each radiologist provided BI-RADS lesion descriptors for the calcifications, BI-RADS final assessment category for the calcifications, and estimated the likelihood that the calcifications were associated with malignancy on a 100-point quasi-continuous scale. The BI-RADS lesion descriptors included descriptors for calcification distribution, calcification morphology, and calcification number. While calcification number is not a BI-RADS descriptor *per se*, it is a descriptor commonly used in computer analysis of radiologists-provided lesion descriptions [5, 6] and we used this descriptor [8] in this study as well. Calcification distribution consisted of 5 different descriptors; calcification morphology consisted of 14 different descriptors; and calcification number consisted of 4 different descriptors: less than 5, 5-10, 10-30, or greater than 30 [7]. Because it was often difficult for the radiologists to decide on a single calcification morphology descriptor for the calcifications, they selected up to two calcification morphology descriptors for each case. When these descriptors were later analyzed with computer classifiers, we needed to input the same number of lesion descriptors to the computer classifier. For the cases that a radiologist provided a single lesion morphology descriptor we duplicated this descriptor and made two lesion morphology descriptors that were identical. Therefore, all cases had two calcification morphology descriptors; in some cases the two descriptors were identical.

2.2. Computer classification of calcifications as malignant or benign

We have developed a computer technique that classifies breast calcifications as malignant or benign based on computer-extracted image features from mammograms [8]. We have shown previously that this computer technique can be as accurate as, or more accurate than, radiologists in classifying calcifications in mammograms, and more importantly, this computer technique can help radiologists improve diagnostic performance in making biopsy recommendations [1, 3, 8]. This computer technique has been described in detail elsewhere [8]. Briefly, the computer extracts eight image features from digital mammograms. These image features describe the size and shape of a calcification cluster, the average and variation in size (including contrast) of individual calcifications, and the degree of shape-irregularity of the individual calcifications. The computer then uses an artificial neural network (ANN) to merge the image features into a single output, and subsequently converts this output to an estimate of the likelihood of malignancy.

2.3. Data analysis

Several different analyses of the data were carried out. The radiologists' BI-RADS final assessments and their estimate of the likelihood of malignancy on the quasi-continuous scale were analyzed with receiver operating characteristic (ROC) analysis directly. Artificial neural networks were developed to analyze radiologists' BI-RADS lesion descriptors and computer-extracted image features in a variety of ways. First, ANNs were developed to analyze radiologists' BI-RADS lesion descriptors alone. Second, an ANN was developed to analyze computer-extracted image features alone.

Third, ANNs were developed to analyze radiologists' BI-RADS lesion descriptors and computer-extracted image features in combination. In these analyses, a single radiologist's BI-RADS data were used for both ANN training and testing. The leave-one-out method was used to avoid training bias. In addition to these, a fourth analysis was carried out with artificial neural networks in which an ANN was trained with one radiologist's BI-RADS data and tested with a different radiologist's BI-RADS data. In this analysis, the ANNs analyzed radiologists' BI-RADS lesion descriptors and computer-extracted image features in combination. The leave-one-out method was also used in this analysis to avoid training bias.

3. RESULTS

The average area under the ROC curve, or A_z value, for BI-RADS final assessments made by the three radiologists was 0.70. The average A_z value for the likelihood of malignancy on a quasi-continuous scale estimated by the three radiologists was 0.75. The standard deviations in both A_z values were 0.05 ($p = 0.049$).

When an artificial neural network was used to analyze BI-RADS lesion descriptors provided by the radiologists, and the same radiologist's data were used for both training and testing, using the leave-one-out method, the average A_z value was 0.71 and the standard deviation was 0.09. When another artificial neural network was used to analyze computer-extracted image features alone, the A_z value was 0.77. When yet another artificial neural network was used to analyze BI-RADS lesion descriptors and computer-extracted image features in combination, still using the same radiologist's data for both training and testing with the leave-one-out method, the average A_z value improved to 0.81 and the standard deviation was 0.04.

When one radiologist's data were used for training while a different radiologist's data were used for testing an artificial neural network, the average A_z values reduced to 0.77 and the standard deviation was 0.05. In this analysis, artificial neural networks were used to analyze BI-RADS data and computer-extracted image features in combination. The leave-one-out method was used in this analysis.

4. DISCUSSION AND SUMMARY

These results indicate that the combination of BI-RADS lesion descriptions provided by radiologists and computer-extracted image features can improve the performance of computer classification of calcifications as malignant or benign. However, variability in radiologists' use of BI-RADS descriptions can diminish that improvement. This finding concerning radiologists' variability in the use of BI-RADS is consistent with other studies of BI-RADS [10, 11], and it raises doubts for the usefulness of BI-RADS lesion descriptors in conjunction with computer-extracted image features for classification of breast lesions as malignant or benign. Such an approach can be expected to work well only if the radiologist uses the BI-RADS descriptors consistently, but it is not clear whether radiologists in general are able to do so at this time. It may be more realistic to expect a radiologist to use the BI-RADS descriptors consistently over time, and one could take advantage of this by developing a computer classifier that is trained on BI-RADS data from, and to be used by, a single radiologist. However, the practical usefulness of such a computer technique would be rather limited.

5. ACKNOWLEDGEMENTS

This work was supported in part by NIH/NCI (R01 CA92361), the US Army Medical Research and Materiel Command (DAMD17-00-1-0197), and Illinois Department of Public Health (Penny Severns Breast and Cervical Cancer Research Fund). RA Schmidt, RM Nishikawa and CJ Vyborny are shareholders of R2 Technology Inc., Sunnyvale, CA.

6. REFERENCES

1. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad Radiol*, vol. 6, pp. 22-33, 1999.
2. H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology*, vol. 212, pp. 817-827, 1999.
3. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, A. Y. Toledano, and K. Doi, "The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms depicting microcalcifications," *Radiology*, vol. 220, pp. 787-794, 2001.
4. Z. Huo, M. L. Giger, C. J. Vybomy, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis -- Observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560-568, 2002.
5. D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," *Invest Radiol*, vol. 23, pp. 240-252, 1988.
6. J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, Jr., "Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology*, vol. 196, pp. 817-822, 1995.
7. American College of Radiology (ACR), *Breast imaging reporting and data system (BI-RADS^{TM})*, Third Edition. Reston, VA: American College of Radiology, 1998.
8. Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vybomy, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, pp. 671-678, 1996.
9. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. J. D'Orsi, C. J. Vybomy, M. L. Giger, L. Lan, Z. Huo, and A. V. Edwards, "Comparison of BI-RADS lesion descriptors and computer-extracted image features for computer classification of malignant and benign breast lesions," in *Digital Mammography 2002*, H. O. Peitgen, Ed. Heidelberg: Springer Verlag Publishers, 2002, pp. 317-321.
10. J. A. Baker, P. J. Kornguth, and C. E. Floyd, Jr., "Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description," *AJR Am J Roentgenol*, vol. 166, pp. 773-778, 1996.
11. W. A. Berg, C. J. D'Orsi, V. P. Jackson, L. W. Bassett, C. A. Beam, R. S. Lewis, and P. E. Crewson, "Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?" *Radiology*, vol. 224, pp. 871-880, 2002.

An ROC comparison of four methods of combining information from multiple images of the same patient

Bei Liu, Charles E. Metz, and Yulei Jiang
Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, Chicago, Illinois 60637

(Received 9 December 2003; revised 12 May 2004; accepted for publication 7 June 2004; published 26 August 2004)

Variance of diagnostic information contained in an image degrades diagnostic accuracy. Acquiring multiple images of the same patient (e.g., mediolateral oblique and craniocaudal view mammograms) can, in principle, help reduce this degradation. We demonstrate how this can be accomplished in the context of computer-aided diagnosis (CAD). Assuming that computer outputs obtained from multiple images of the same patient can be transformed monotonically to the same pair of truth-conditional normal distributions and, for simplicity, ignoring correlation among images, we investigate theoretically four methods of combining the computer outputs: taking the average, the median, the maximum, or the minimum. We found, as one would expect, that both the average and the median always produce an improved area under the receiver operating characteristic (ROC) curve (AUC) compared to the single-view images, while the average always produces better performance than the median. However, the maximum and minimum also can produce improved AUCs in some situations, and under certain conditions can outperform the average. Surprisingly, we found that the maximum and minimum of normally-distributed decision variables produce nearly binormal ROC curves. These results can be used as a guide in attempting to increase the efficacy of CAD when multiple images are available from the same patient. © *2004 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1776674]

Key words: ROC analysis, binormal model, computer-aided diagnosis, multiple medical images, combination of multiple assessments

I. INTRODUCTION

Variance in diagnostic information degrades the accuracy of diagnosis by broadening the statistical distributions of diagnostic information and thus blurring the distinction between diagnostic information from healthy and diseased patients.[1] Metz and Shen investigated the degradation of diagnostic accuracy introduced by variance in human interpretation of medical images and calculated the gain in diagnostic accuracy that is available from averaging repeated readings of the same images by accuracy-equivalent readers.[2] Swensson *et al.* showed that the median can be used to achieve a similar gain in diagnostic accuracy.[3]

Acquiring multiple images can reduce the degradation of diagnostic accuracy due to the variance of diagnostic image information and is widely believed to help improve the accuracy of diagnosis. It is standard practice in mammography to acquire two images, the mediolateral oblique (MLO) and craniocaudal (CC) views. This presents a dilemma for computer-aided diagnosis (CAD), however. On one hand, CAD can take a patient-based approach: to treat all images of a patient as a unit and analyze these images as a whole. The easiest way to deal with features from different images (e.g., lesion area in an MLO view mammogram and the same feature in a CC view mammogram) is to treat them as if they were different features. The disadvantage of this approach is that it increases the dimensionality of the analysis, which equals the number of images per patient times the number of image features extracted from each image. Increased dimensionality makes it more difficult to obtain a reliable computer classifier. On the other hand, CAD can take an image-based approach, treating each image as a unit and analyzing multiple images of the same patient independently. However, the results of this analysis of multiple images of a patient must be combined into a single diagnostic variable, and the optimal method for doing so is unknown.

Researchers have employed ad hoc methods to combine results of computer analyses of multiple images of the same patient. Jiang *et al.* used the maximum output, which corresponds to the highest likelihood of malignancy in classifying malignant and benign microcalcifications in mammograms.[4] Chan *et al.* compared the use of average and maximum in calculating the likelihood of malignancy of microcalcifications and found the method of average to be slightly better.[5] Huo *et al.* compared the average, maximum and minimum in classifying malignant and benign masses and found average to be the best.[6] However, to our knowledge, no one has reported a general approach to the question of which method produces the best result in which situations.

We have studied theoretically the diagnostic performance obtained from *average, median, maximum* and *minimum* values within the framework of ROC analysis. For clarity, we use italics in this paper to indicate a particular method of combining multiple diagnostic assessments: the average, the median, the maximum, and the minimum. We also refer to the use of a single output per patient as the *single-output* method. To simplify the analysis, we assumed in this work that each classifier output (e.g., from the MLO or CC view) produces the same ROC curve when employed alone. While this work was motivated by CAD of multiple images from the same patient, the question applies in principle to any decision task requiring the combination of multiple assessments for each case. Thus, to be general, we will use the term "per case" instead of "per patient" hereafter in this article. In Sec. II A, we review the binormal model for ROC analysis, upon which our theory is based. In Sec. II B, we derive the theoretical relationship between true positive fraction (TPF) and false positive fraction (FPF) for each method. In Sec. III, we calculate areas under ROC curves (AUCs) and identify the situations in which each method produces the best result. We then provide in Sec. IV an intuitive explanation for our results and discuss limitations and potential applications of the theory, followed by conclusions in Sec. V.

II. THEORY

A. The binormal model

ROC analysis is widely recognized as the most effective and meaningful way to quantify diagnostic performance of an imaging systems.[1,7] An ROC curve completely describes all available tradeoffs between sensitivity and specificity. Sensitivity or TPF is defined as the probability that an actually positive case will be diagnosed as positive. Specificity is defined as the probability that an actually negative case will be diagnosed as negative. Instead of specificity, it is often more convenient to use FPF, which is defined as the probability that an actually negative case will be diagnosed as positive and equals 1-specificity. Although an entire ROC curve is required to describe the performance of an imaging system completely, the area under the ROC curve (AUC) is a useful summary index that can be interpreted as the average value of sensitivity over all possible values of specificity or, equivalently, as the average value of specificity over all possible values of sensitivity.

In ROC analysis, the binormal model has been used successfully to fit data obtained in a wide variety of practical situations.[8,9] This model assumes that the observed decision variable can be transformed to a pair of normal (i.e., Gaussian) distributions by an unknown monotonic transformation of the decision variable, whence:

$$f(x|\text{negative}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \tag{1}$$

for actually negative cases, and

$$f(x|\text{positive}) = \frac{b}{\sqrt{2\pi}} \exp\left\{-\frac{(bx-a)^2}{2}\right\} \tag{2}$$

for actually positive cases, in which x is a latent decision variable. In this article, we use the notation $N(\mu,\sigma)$ to describe a normal distribution with mean μ and standard deviation σ. With this notation, the binormal model can be expressed as $N(0,1)$ for actually negative cases and $N(a/b,1/b)$ for actually positive cases.

Given the binormal model, it is straightforward to calculate TPF, FPF, and AUC. With a critical value x_c of the latent decision variable x, a case is diagnosed as positive if and only if $x > x_c$. Therefore, TPF(x_c) is given by

$$\text{TPF}(x_c) = \int_{x_c}^{\infty} \frac{b}{\sqrt{2\pi}} \exp\left\{-\frac{(bx-a)^2}{2}\right\} dx = \Phi(a - bx_c) \tag{3}$$

and FPF(x_c) is given by

$$\text{FPF}(x_c) = \int_{x_c}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = \Phi(-x_c), \tag{4}$$

in which $\Phi(z)$ represents the cumulative standard normal distribution function. As x_c varies from $-\infty$ to ∞, TPF(x_c) and FPF(x_c) sweep out the ROC curve, which can be represented in a unit square by plotting TPF as a function of FPF. The area under this binormal ROC curve, denoted by A_z with the subscript z indicating use of the binormal model, can be expressed as (see Appendix A for derivation)[2,10]

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{5}$$

Alternatively, the same ROC curve can be plotted on "normal-deviate axes," Z_{TPF} as a function of Z_{FPF}, where $\text{TPF} = \Phi(Z_{\text{TPF}})$ and $\text{FPF} = \Phi(Z_{\text{FPF}})$. Applying the inverse cumulative standard normal distribution function to both sides of Eqs. (3) and (4), we have

$$Z_{\text{TPF}} = a - bx_c \tag{6}$$

and

$$Z_{\text{FPF}} = -x_c, \tag{7}$$

whence

$$Z_{\text{TPF}} = a + bZ_{\text{FPF}}. \tag{8}$$

Thus, a binormal ROC curve plots on normal-deviate axes as a straight line with slope b and "y-intercept" a.[1,7] Because any pair of distributions that can be transformed monotonically to the binormal distributions $N(0,1)$ and $N(a/b,1/b)$ will have the same ROC curve,[9] the extent to which the ROC curve obtained from any arbitrary pair of decision-variable distributions is approximated well by a straight line on normal-deviate axes indicates how closely the underlying distributions can be transformed monotonically to a pair of normal distributions.

B. Combining information derived from multiple images of the same patient

Acquiring multiple images of the same patient can increase diagnostic accuracy. In many CAD approaches, multiple images of the same patient produce multiple diagnostic outputs. To simplify the analysis, we assume that each individual output yields the same ROC curve. We now analytically derive the ROC curves produced by several methods—the *average*, the *median*, the *maximum*, and the *minimum*—for combining these multiple outputs into one decision variable. To the extent that each output's ROC curve plots as a straight line on normal deviate axes, one can always, in principle, transform the individual outputs monotonically to binormal distributions. Therefore, we can, without loss of generality, start from the same binormal distributions for each individual outputs: $N(0,1)$ and $N(a/b,1/b)$, whose *single-output* A_z value is $\Phi(a/\sqrt{1+b^2})$. We further assume that the monotonic transformation which connects an observed distribution of the data to the binormal model is known. In practice, this transformation can be estimated by use of the LABROC4 software developed by Metz *et al.*[11] Moreover, for simplicity, we assume that these multiple outputs have no correlation. Although correlations are present in most real-world applications, they substantially complicate the present analysis, and we will defer the consideration of them to future work.

1. *Average*

The average of n normally distributed, independent random variables with mean μ and standard deviation σ is also normally distributed with the same mean but with a standard deviation $\sigma/\sqrt{n}$.[12] Therefore, the result of averaging also follows a pair of normal distributions: $N(0,1/\sqrt{n})$ for actually negative cases and $N(a/b,1/(b\sqrt{n}))$ for actually positive cases. These two normal distributions can be easily transformed to the binormal model: $N(0,1)$ and $N(\sqrt{n}a/b,1/b)$. Thus,

$$A_z=\Phi\left(\frac{a_{\text{avg}}}{\sqrt{1+b_{\text{avg}}^2}}\right)=\Phi\left(\frac{\sqrt{n}a}{\sqrt{1+b^2}}\right). \tag{9}$$

This ROC area is greater than the *single-output* A_z, because the *average* reduces the standard deviation of the normal distributions, thereby reducing the overlap between the actually-negative and actually-positive distributions. On normal-deviate axes, this ROC curve is given by

$$Z_{\text{TPF}}=\sqrt{n}a+bZ_{\text{FPF}}, \tag{10}$$

which has the same slope as the *single-output* ROC curve but a greater y-intercept. As the number of multiple outputs, n, increases, both the y-intercept and the A_z value increase.

2. *Maximum*

Consider the maximum of n normal random variables $x_i\,(i=1,\ldots,n)$ drawn independently from $N(\mu,\sigma)$:

$$x_{\max}=\max\{x_1,\ldots,x_i,\ldots,x_n\}. \tag{11}$$

The probability that $x_{\max}\leq x$, for an arbitrary x, is given by

$$\begin{aligned}P(x_{\max}\leq x)&=P(x_1\leq x,\ldots,\ \text{and}\ x_i\leq x,\ldots,\ \text{and}\ x_n\leq x)\\&=\prod_{i=1}^{n}P(x_i\leq x)=[P(x_1\leq x)]^n\\&=\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^n.\end{aligned} \tag{12}$$

With the binormal model, $\mu=0$ and $\sigma=1$ for actually negative cases, so

$$P(x_{\max}\leq x|\text{negative})=[\Phi(x)]^n. \tag{13}$$

Similarly, $\mu=a/b$ and $\sigma=1/b$ for actually positive cases, whence:

$$P(x_{\max}\leq x|\text{positive})=\left[\Phi\left(\frac{x-a/b}{1/b}\right)\right]^n=[\Phi(bx-a)]^n. \tag{14}$$

Comparing $x_{\max}$ to a critical value, x_c, to calculate TPF and FPF, we obtain

$$\begin{aligned}\text{TPF}(x_c)&=P(x_{\max}>x_c|\text{positive})\\&=1-P(x_{\max}\leq x_c|\text{positive})=1-[\Phi(bx_c-a)]^n\end{aligned} \tag{15}$$

and

$$\begin{aligned}\text{FPF}(x_c)&=P(x_{\max}>x_c|\text{negative})\\&=1-P(x_{\max}\leq x_c|\text{negative})=1-[\Phi(x_c)]^n.\end{aligned} \tag{16}$$

From Eq. (16),

$$x_c=\Phi^{-1}(\sqrt[n]{1-\text{FPF}}), \tag{17}$$

so from Eq. (15), we have

$$\text{TPF}=1-[\Phi(b\Phi^{-1}(\sqrt[n]{1-\text{FPF}})-a)]^n \tag{18}$$

and

$$Z_{\text{TPF}}=\Phi^{-1}\{1-[\Phi(b\Phi^{-1}(\sqrt[n]{\Phi(-Z_{\text{FPF}})})-a)]^n\}. \tag{19}$$

From Eq. (18) we can calculate the AUC of the *maximum* by numerically integrating the expression $\int_0^1\text{TPF(FPF)}\,d(\text{FPF})$, whereas from Eq. (19) we can evaluate numerically the linearity of the ROC curve on normal-deviate axes: Z_{TPF} versus Z_{FPF}.

3. *Minimum*

The functional relationship between Z_{TPF} and Z_{FPF} for the *minimum* can be derived by following a procedure similar to that used for the *maximum.* Consider the minimum of n normally distributed, independent random variables x_i $(i=1,\ldots,n)$ drawn from $N(\mu,\sigma)$:

$$x_{\min}=\min\{x_1,\ldots,x_i,\ldots,x_n\}. \tag{20}$$

The probability for $x_{\min}>x$, of any arbitrary x, is given by

$$\begin{aligned}P(x_{\min}>x)&=P(x_1>x,\ldots,\ \text{and}\ x_i>x,\ldots,\ \text{and}\ x_n>x)\\&=\prod_{i=1}^{n}P(x_i>x)=[P(x_1>x)]^n\\&=\left[\Phi\left(-\frac{x-\mu}{\sigma}\right)\right]^n.\end{aligned} \tag{21}$$

Therefore, for a critical value x_c of the decision variable, TPF and FPF of the *minimum* are given by

$$\text{TPF}(x_c)=P(x_{\min}>x_c|\text{positive})=[\Phi(a-bx_c)]^n \tag{22}$$

and

$$\text{FPF}(x_c)=P(x_{\min}>x_c|\text{negative})=[\Phi(-x_c)]^n. \tag{23}$$

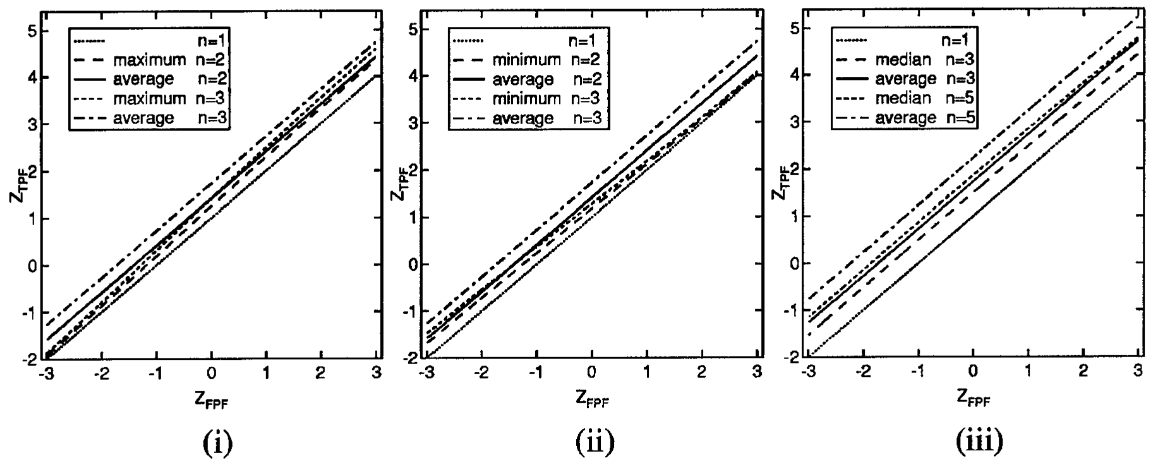

FIG. 1. Normal deviate ROC curves derived from single-output binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$: (i) ROC curves of the *maximum* and the *average* for $n=1$, 2 and 3; (ii) ROC curves of the *minimum* and the *average* for $n=1$, 2 and 3; (iii) ROC curves of the *median* and the *average* for $n=1$, 3 and 5. For $n=1$, the ROC curves of the *maximum*, the *minimum*, the *average* and *median* are the same, which is the *single-output* ROC curve, so we did not label $n=1$ in the figure.

From Eq. (23), we have

$$x_c = -\Phi^{-1}(\sqrt[n]{\mathrm{FPF}}). \tag{24}$$

Plug Eq. (24) into Eq. (22), we have

$$\mathrm{TPF} = [\Phi(a + b\Phi^{-1}(\sqrt[n]{\mathrm{FPF}}))]^n, \tag{25}$$

so

$$Z_{\mathrm{TPF}} = \Phi^{-1}\{[\Phi(a + b\Phi^{-1}(\sqrt[n]{\Phi(Z_{\mathrm{FPF}})}))]^n\}. \tag{26}$$

Thus, we can calculate the AUC of the *minimum* numerically from Eq. (25) and evaluate numerically the linearity of the ROC curve on normal-deviate axes from Eq. (26).

4. Median

The median of a set of variables is found by arranging their values in rank order and then selecting the one in the middle. If the number of variables is even, then the median is defined as the average of the two numbers in the middle. If there are just two variables, then the median is simply the average. We now derive the ROC curve obtained from the median of an odd number of outputs: $n = 2m+1$, where m is a natural number.

Given a critical value x_c of the decision variable, the probability that a single output of an actually positive case yields a positive diagnosis is given by

$$p_p \equiv P(x > x_c|\text{positive}) = \Phi(a - bx_c), \tag{27}$$

whereas the probability that a single output of an actually negative case yields a positive diagnosis is given by

$$p_n \equiv P(x > x_c|\text{negative}) = \Phi(-x_c). \tag{28}$$

With $2m+1$ independent outputs, the median of these outputs will yield a positive diagnosis if and only if at least $m+1$ of these outputs indicate a positive diagnosis. Therefore, TPF for the *median* can be written as

$$\begin{aligned}\mathrm{TPF}(x_c) &= \sum_{k=m+1}^{n=2m+1} \binom{n}{k} p_p^k (1-p_p)^{n-k} \\ &= \sum_{k=m+1}^{n=2m+1} \binom{n}{k} [\Phi(a - bx_c)]^k \\ &\quad \times [1 - \Phi(a - bx_c)]^{n-k},\end{aligned} \tag{29}$$

where $\binom{k}{n} \equiv n!/k!(n-k)!$. Similarly, FPF for the *median* can be written as

$$\mathrm{FPF}(x_c) = \sum_{k=m+1}^{n=2m+1} \binom{n}{k} [\Phi(-x_c)]^k [1 - \Phi(-x_c)]^{n-k}. \tag{30}$$

The quantities $Z_{\mathrm{TPF}}(x_c)$ and $Z_{\mathrm{FPF}}(x_c)$ can be obtained by applying the inverse cumulative standard normal distribution function to both sides of Eqs. (29) and (30). Note that unlike the *average, maximum*, and *minimum*, the functional relationships between TPF and FPF and between Z_{TPF} and Z_{FPF} for the *median* must be described implicitly through the critical value x_c. However, the AUC of the *median* can still be calculated numerically by integrating $\int_{-\infty}^{\infty} -\mathrm{TPF}(x_c) \times [d\mathrm{FPF}(x_c)/dx_c]dx_c$, and the linearity of the ROC curve on normal-deviate axes that is obtained from the *median* can be investigated numerically by calculating Z_{TPF} and Z_{FPF} pairs from a variety of x_c values.

III. RESULTS

A. Linearity of ROC curves on normal-deviate axes

The binormal ROC curve given by $N(0,1)$ and $N(a/b,1/b)$ plots as a straight line on normal-deviate axes with slope b and y-intercept a. According to Eq. (10), the *average* produces strictly binormal results, so its ROC curve also plots as a straight line on normal-deviate axes. According to Eqs. (19), (26), (29), and (30), the *maximum, mini-*

TABLE I. Eighth order polynomial fit of normal-deviate ROC curves of the *maximum*, the *minimum*, the *average*, and the *median*, derived from single-output binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$ with $a=1$ and $b=1$. The coefficients of determination (R^2) are essentially equal to 1 in all fits. P_i denotes the fitted coefficient of the ith order term.

Method	N, number of multiple outputs	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
Maximum	2	1.2518	1.0509	−0.0079	−0.0002	0.0005	−0.0001	0.0000	0.0000	0.0000
	3	1.4147	1.0902	−0.0123	−0.0007	0.0008	−0.0001	0.0000	0.0000	0.0000
	4	1.5371	1.1225	−0.0151	−0.0013	0.0011	−0.0002	0.0000	0.0000	0.0000
	5	1.6359	1.1501	−0.0171	−0.0019	0.0013	−0.0002	0.0000	0.0000	0.0000
Minimum	2	1.1819	0.9389	−0.0018	0.0026	0.0004	−0.0001	0.0000	0.0000	0.0000
	3	1.2828	0.9008	0.0002	0.0038	0.0004	−0.0002	0.0000	0.0000	0.0000
	4	1.3512	0.8739	0.0025	0.0045	0.0003	−0.0002	0.0000	0.0000	0.0000
	5	1.4023	0.8535	0.0046	0.0049	0.0002	−0.0002	0.0000	0.0000	0.0000
Average	2	1.4142	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.7321	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	2.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	2.2361	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Median	3	1.4945	0.9893	−0.0065	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000
	5	1.8658	0.9857	−0.0070	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	7	2.1755	0.9840	−0.0067	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000
	9	2.4465	0.9830	−0.0064	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000

mum, and *median* do not produce strictly binormal results; however, they produce nearly linear ROC curves on normal-deviate axes, as we show below.

Figure 1 compares the ROC curves produced by the *maximum*, the *minimum* and the *median* to that produced by the *average* on normal-deviate axes for $a=1$ and $b=1$. The range of Z_{FPF} shown in this figure, −3.0 to 3.0, corresponds to FPF values from 0.001 to 0.999. The ROC curves obtained from the *average* of n independent decision variables are parallel straight lines with a y-intercept that increases as n increases. Perhaps surprisingly, however, the ROC curves of the *maximum*, the *minimum*, and the *median* also are nearly linear on normal-deviate axes. The slope of the ROC curve from the *maximum* increases slightly as n increases and is greater than the slope of the corresponding ROC curve for the *average*, whereas the opposite is true for the *minimum*, which yields a slope that decreases slightly as n increases and is smaller than the slope of the corresponding ROC curve for the *average*. The slopes of the *median* are similar to those of the *average*, an observation that can be ascribed to the fact that both median and average are location measures.

We performed linear regression analysis on these ROC curves. The coefficient of determination[13] was found to be $R^2 \simeq 0.9999$ in all such analyses. The nonlinear coefficients (Table I) are at least several orders of magnitude smaller than the linear coefficient, indicating quantitatively that the ROC curves are virtually straight lines. The linearity of the *maximum* and *minimum* ROC curves decreases slightly as n increases, whereas the linearity of the *median* is nearly independent of n (Fig. 2).

B. Comparison of AUCs

Figure 3 compares the *single-output* ROC curve given by the binormal model $N(0,1)$ and $N(a/b,1/b)$ with $a=1$ and $b=1$, and corresponding ROC curves obtained from the *average*, the *median*, the *maximum*, and the *minimum* for $n=2$ and $n=3$. The A_z value of the *single-output* ROC curve is 0.760, given by Eq. (5). The A_z values of the *average* ROC curves are 0.841 for $n=2$ and 0.890 for $n=3$, from Eq.

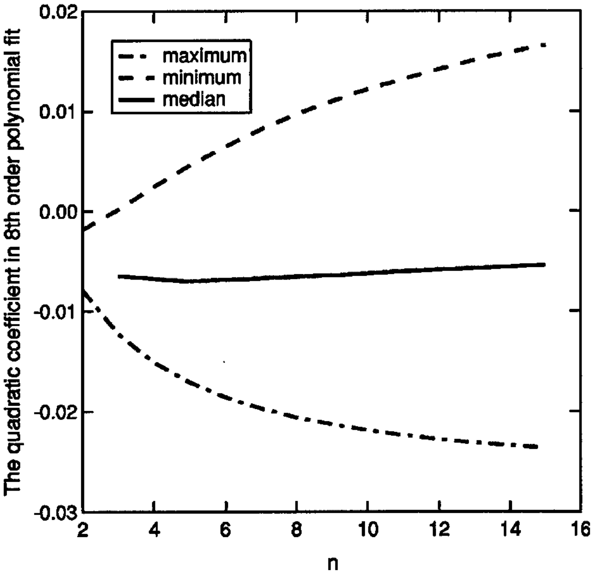

FIG. 2. The quadratic coefficient (P2 in Table I) in 8th order polynomial fit of the normal deviate ROC curves of the *maximum*, the *minimum* and the *median* of n outputs.

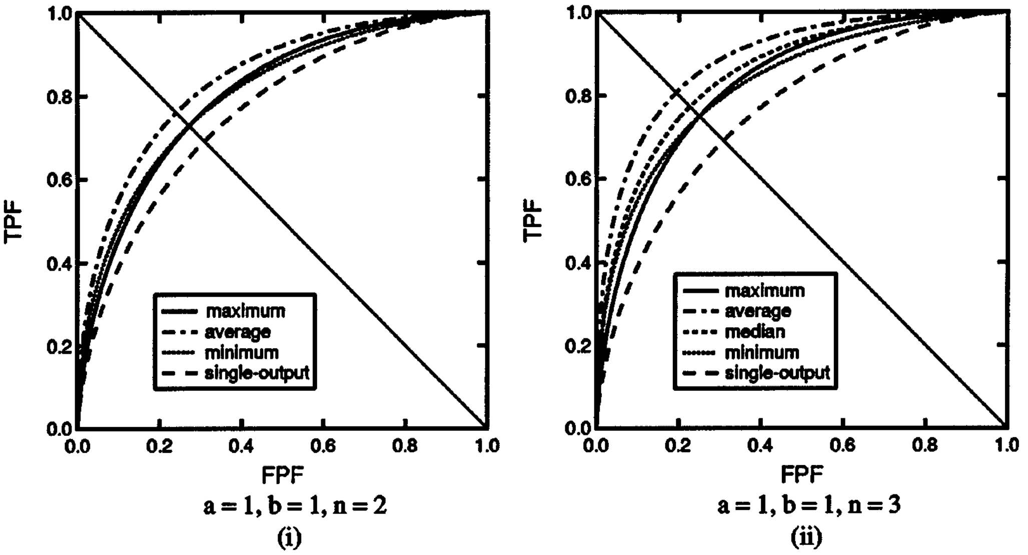

FIG. 3. ROC curves of the *average*, the *median*, the *maximum* and the *minimum* plotted on probability axes for $n=2$ (i) and 3 (ii), derived from the *single-output* given by the binormal distribution pair N (0,1) and N $(a/b,1/b)$, with $a=1$ and $b=1$. The negative diagonal of the unit square is shown as a solid line in order to demonstrate the skewness of the *maximum* and the *minimum*.

(9). The AUCs for the *maximum, minimum*, and *median* ROC curves were calculated from Eqs. (18), (25), (29), and (30) by numerical integration: AUC=$\int_0^1$TPF(FPF) d(FPF). The AUC for the *median* ROC curve is 0.855 for $n=3$. The AUC values for the *maximum* and *minimum* ROC curves are the same: 0.805 for $n=2$ and 0.829 for $n=3$; however, the *maximum* ROC curve is slightly higher at high sensitivities, whereas the *minimum* ROC curve is slightly higher at low sensitivities. The pair of ROC curves obtained from the *maximum* and the *minimum* are skewed equally in opposite directions around the negative diagonal of the unit square because, for $b=1$, the two distributions of the binormal model are symmetric, whereas the *maximum* and the *minimum* are equivalent under a reflection of the decision-variable axis.

Figure 4 compares the AUCs of the *single-output* ROC curves and of the ROC curves from the *average*, the *median*, the *maximum*, and the *minimum*, with $a=1$, $n=2$ and with $a=1$, $n=3$ for various values of b, which represents the relative widths of the two truth-conditional normal distributions for a single output. As one would expect, the A_z value of the *average* is greater than that of the *single-output* ROC curve, because the *average* does not change the difference between the means of the distributions of negative and positive cases but narrows both distributions equally, thereby reducing the overlap between negative and positive distributions and hence increasing the A_z value. The *median*, a location measure like the *average*, also always improves the AUC, but its performance is never as good as the *average*. Therefore, we will not consider *median* further in this article.

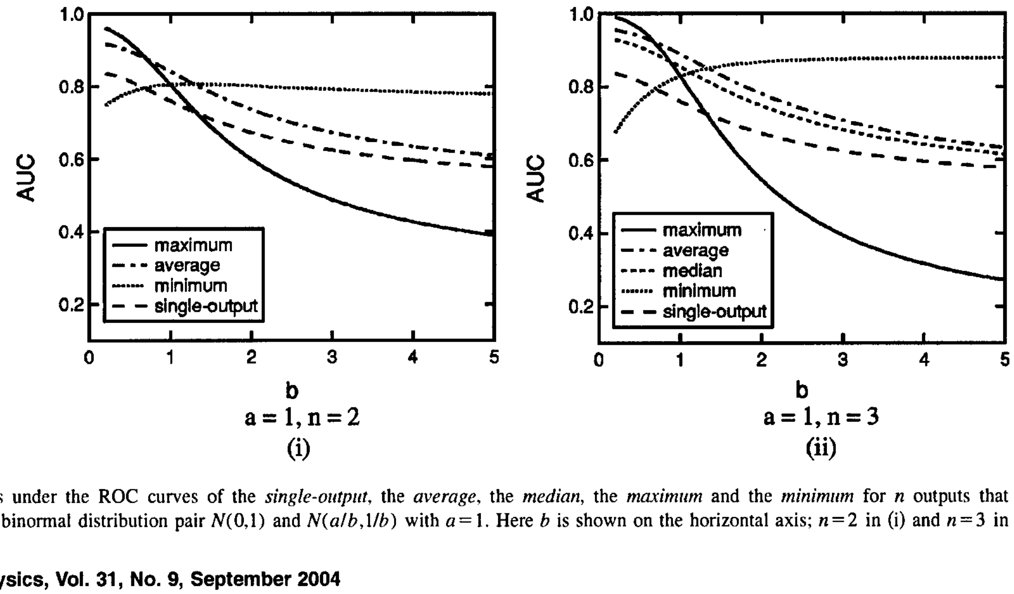

FIG. 4. Areas under the ROC curves of the *single-output*, the *average*, the *median*, the *maximum* and the *minimum* for n outputs that arise from the single-output binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$ with $a=1$. Here b is shown on the horizontal axis; $n=2$ in (i) and $n=3$ in (ii).

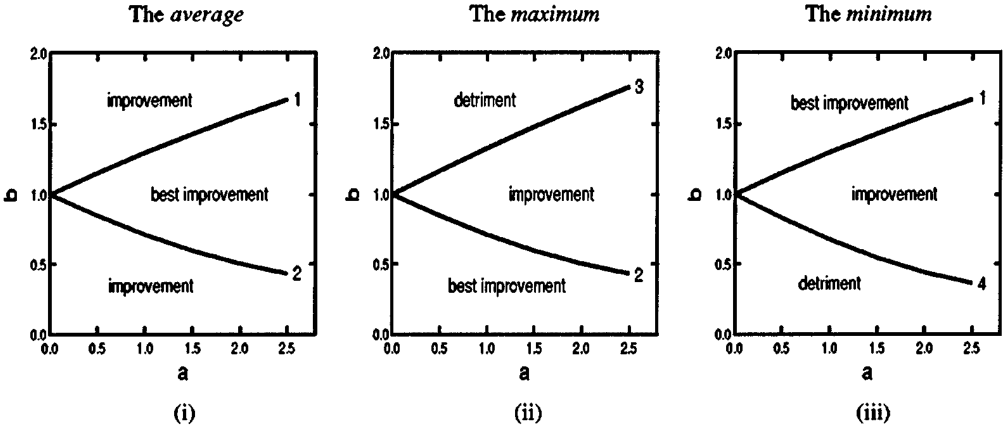

FIG. 5. Curves showing combinations of parameters a and b such that two methods among the *average*, the *maximum*, the *minimum* and the *single-output* produce the same AUC for $n=2$. These curves partition the (a,b) space into three labeled regions for each method according to their performance levels. See text for definition of the labeling of the curves and regions of (a,b) space.

In contrast, the *maximum* and the *minimum* improve AUC in some situations, and outperform the *average* under certain conditions. However, the *maximum* and the *minimum* can also produce AUC values smaller than that of the *single-output*. Any intersection between two curves in Fig. 4 indicates that the two corresponding methods produce the same AUC result. The curves for the *average* and *median* do not intersect that for the *single-output* because these methods always produce better results. All other curves intersect each other.

Figure 5 compares the relative performance of the *average*, the *maximum*, and the *minimum* in terms of the AUCs that these methods produce in the plane of a and b, which are the binormal model parameters of the *single-output* ROC curve. In this figure, the *average* and the *minimum* produce the same AUC along curve 1; the *average* and the *maximum* produce the same AUC along curve 2; the *maximum* and the *single-output* produce the same AUC along curve 3; and the *minimum* and *single-output* produce the same AUC along curve 4. The *maximum* and the *minimum* produce the same AUC along the straight line $b=1$, which is not shown in this plot. These curves partition the (a,b) space into three labeled regions for each method, where the region labeled "best improvement" indicates that particular method produces the best AUC among the methods of the *average*, the *maximum* and the *minimum*; the region labeled with "improvement" indicates better AUC than that of the *single-output* but not the best AUC; and the region labeled "detriment" indicates lower AUC than that of the *single-output*. Figure 5(i) identifies combinations of a and b such that the *average* produces the best AUC, and combinations of a and b such that the average produces an improved, but not the best, AUC. Similarly, Figs. 5(ii), and 5(iii) identify such combinations of a and b for the *maximum* and the *minimum*, and additionally, identify combinations of a and b such that these methods produce detrimental results, i.e., AUC values smaller than that of the *single-output*.

Figure 6 shows how the regions of best performance of each method change with n, the number of multiple outputs for each case. As n increases, the region in which the *average* (region I) produces the best performance becomes larger, while the regions in which the *maximum* (region II) and the *minimum* (region III) produce the best performance become smaller.

IV. DISCUSSION

A. An intuitive explanation for improved ROC curves

In Sec. II we derived analytically the functional relationship between TPF and FPF for the *average, maximum,* and *minimum* methods. AUC calculations show that although the *average* always improves performance, the *maximum* or the

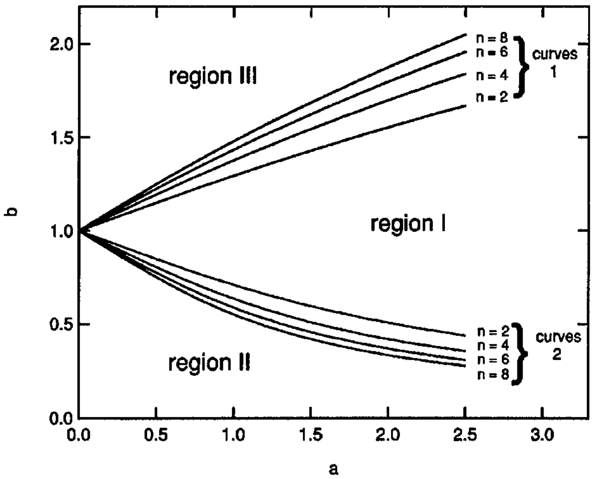

FIG. 6. The best performance regions of the *average*, the *maximum* and the *minimum* versus n, the number of multiple outputs for each case. The definitions of curves 1 and 2 are the same as in Fig. 5. Region II and region III are the regions below curve 2 and above curve 1, respectively. Region I is the region between curve 1 and curve 2.

minimum can perform even better under certain conditions. In this section, we provide an intuitive explanation for these results.

The maximum of n normally distributed, independent random variables, $x_i\,(i=1,2,...,n)$, drawn from $N(\mu,\sigma)$ satisfies Eq. (12); therefore, the probability density function of $x_{\max} \equiv x$ is

$$f_{\max}(x_{\max}=x)=\frac{d[P(x_{\max}\leqslant x)]}{dx}$$

$$=\frac{d\left(\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{n}\right)}{dx}$$

$$=n\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{n-1} f(x|\mu,\sigma), \qquad (31)$$

where $f(x|\mu,\sigma)$ denotes the probability density of $N(\mu,\sigma)$. Similarly, According to Eq. (21), the probability of $x_{\min}\leqslant x$ is

$$P(x_{\min}\leqslant x)=1-\left[\Phi\left(-\frac{x-\mu}{\sigma}\right)\right]^{n}. \qquad (32)$$

Therefore, the probability density of $x_{\min}$ is

$$f_{\min}(x_{\min}=x)=\frac{d[P(x_{\min}\leqslant x)]}{dx}$$

$$=n\left[\Phi\left(-\frac{x-\mu}{\sigma}\right)\right]^{n-1} f(x|\mu,\sigma). \qquad (33)$$

From Eqs. (31) to (33), we can calculate the mean and standard deviation of $x_{\max}$ and $x_{\min}$. Appendix B derives these for the special case of $n=2$: the mean and standard deviation of $x_{\max}$ and $x_{\min}$ are given in Eqs. (B5)–(B8). One can see that the standard deviations of $x_{\max}$ and $x_{\min}$ become smaller, the mean of the $x_{\max}$ becomes greater, and the mean of $x_{\min}$ becomes smaller than the corresponding parameters of the original *single-output* distributions. It is straightforward to calculate the mean and standard deviation of the average; for $n=2$ in particular

$$\overline{x_{\text{avg}}}=\mu \qquad (34)$$

and

$$\text{SD}_{x_{\text{avg}}}=\frac{\sigma}{\sqrt{2}}. \qquad (35)$$

Clearly, narrowing the distribution width and/or increasing the separation between the two distribution means will improve the ROC curves. All three methods narrow the distribution width: the *average* reduces σ to $\sigma/\sqrt{2}=0.707\sigma$, whereas the *maximum* and *minimum* reduce σ to $\sqrt{1-1/\pi}\sigma=0.826\sigma$. These three methods have different effects on the means of the distributions, however. The *average* does not change the means of the distributions, whereas the *maximum* moves the means of the distributions to the right and the *minimum* moves the means of the distributions to the left, with the amount of the movement proportional to standard deviation of the original distributions. Therefore, the *average* always improves the ROC curve, whereas the effects of the *maximum* and *minimum* will depend on the relative and sometimes competing effects of distribution width and separation.

For a *single-output* binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$, the difference between the means for positive cases and negative cases is $a/b-(b-1)/(b\sqrt{\pi})$ for the *maximum* and $a/b+(b-1)/(b\sqrt{\pi})$ for the *minimum.* These two quantities represent a widening of the separation between the means of the two distributions when $b<1$ for the *maximum* and when $b>1$ for the *minimum.* Therefore, the *maximum* for $b<1$ and the *minimum* for $b>1$ will improve the ROC curve; they can even outperform the *average* because they reduce distribution width and widen the separation between their means at the same time. On the other hand, the *minimum* for $b<1$ and the *maximum* for $b>1$ can lower the ROC curve when their effects on reducing the separation between the means of the distributions surpass their effect on narrowing the distribution widths.

Figure 7 shows three sets of ROC curves corresponding to the binormal distribution pairs ($a=1$, $b=1$), ($a=1$, $b=2$), and ($a=1$, $b=0.5$), along with probability-density pairs obtained from the *average*, the *maximum*, and the *minimum.* One can see, at least qualitatively, that the *average* always reduces the overlap between the two distributions, whereas the *maximum* and the *minimum* have varied effects on the overlap. For ($a=1$, $b=1$), the *average* produces the smallest overlap, whereas the *maximum* and *minimum* produce smaller and almost equal overlaps, indicating best ROC curve from the *average* and improved ROC curves from the *maximum* and *minimum.* For ($a=1$, $b=2$), however, the *minimum* produces the smallest overlap, indicating that the best ROC curve is obtained from the *minimum*, whereas for ($a=1$, $b=0.5$) the *maximum* produces the smallest overlap, indicating the best ROC curve from the *maximum.*

B. Gaussian approximation to probability densities of maximum and minimum

The observation that ROC curves produced by the *maximum* and *minimum* methods plot as nearly straight lines on normal-deviate axes (Fig. 1) implies that the distributions of the *maximum* and the *minimum* can be transformed monotonically to distributions that are very close to normal. In this section we use Gaussian functions to approximate the probability density functions given by Eqs. (31) and (33), which are not exactly Gaussian, and calculate the AUCs of the *maximum* and the *minimum* for $n=2$. According to Eqs. (B5)–(B8), we approximate the probability densities of the *maximum* by two Gaussian distributions: $N(1/\sqrt{\pi},\sqrt{1-1/\pi})$ and $N((a/b)+(1/b\sqrt{\pi}),(1/b)\sqrt{1-1/\pi})$, and we approximate the probability densities of the *minimum* by $N(-1/\sqrt{\pi},\sqrt{1-1/\pi})$ and $N((a/b)-(1/b\sqrt{\pi}),(1/b)\sqrt{1-1/\pi})$. For the *maximum* and the *minimum*, the ratios of standard deviations of the actually-negative distribution to the actually-positive distribution are still b, and the differences between the means of the actually-negative distribution and the

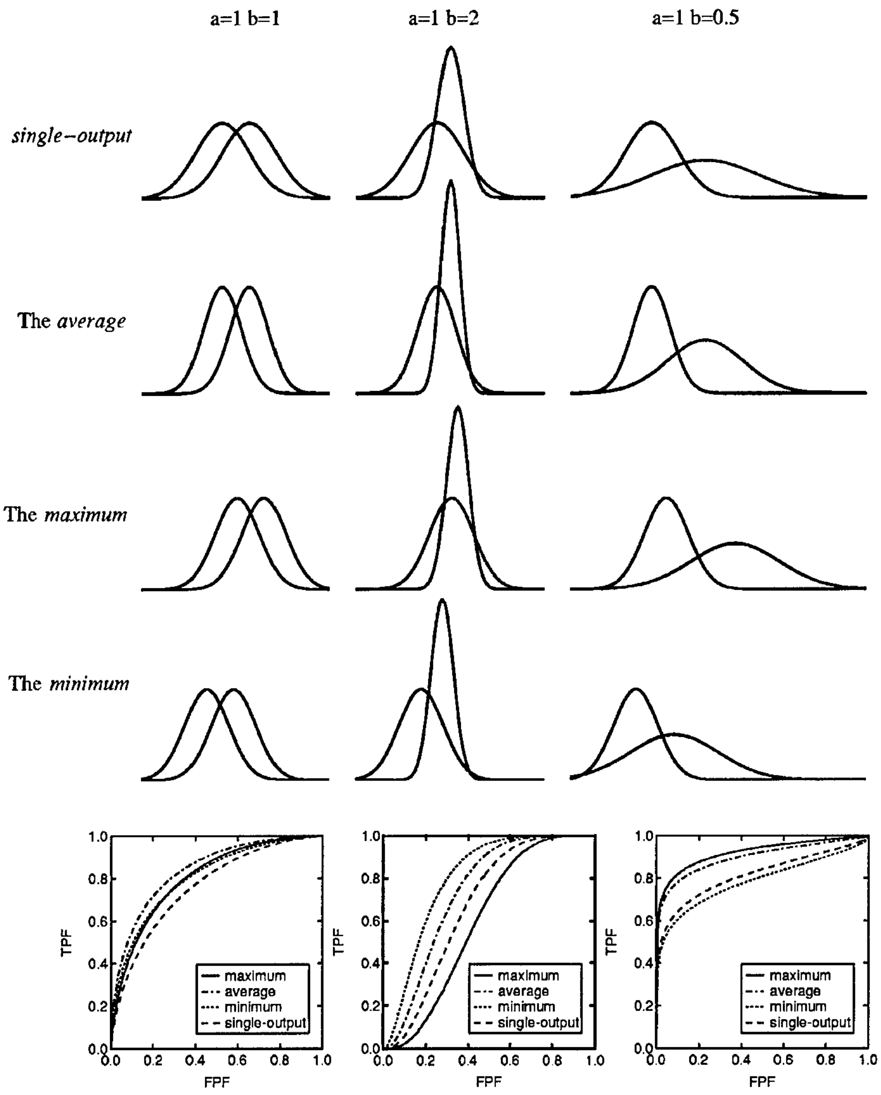

FIG. 7. *Single-output* binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$ for $a=1$, $b=1$, for $a=1$, $b=2$ and for $a=1$, $b=0.5$, together with the probability densities obtained from the *average*, the *maximum* and the *minimum* for $n=2$. The corresponding ROC curves are also shown.

actually-positive distribution expressed in units of the standard deviation of the actually-positive distribution are $(\sqrt{\pi}a-b+1)/\sqrt{\pi-1}$ and $(\sqrt{\pi}a+b-1)/\sqrt{\pi-1}$, respectively. Therefore[2]

$$\mathrm{AUC}_{\max}=\Phi\left(\frac{\frac{\sqrt{\pi}a-b+1}{\sqrt{\pi-1}}}{\sqrt{1+b^2}}\right)=\Phi\left(\frac{\sqrt{\pi}a-b+1}{\sqrt{(\pi-1)(1+b^2)}}\right) \tag{36}$$

and

$$\mathrm{AUC}_{\min}=\Phi\left(\frac{\frac{\sqrt{\pi}a+b-1}{\sqrt{\pi-1}}}{\sqrt{1+b^2}}\right)=\Phi\left(\frac{\sqrt{\pi}a+b-1}{\sqrt{(\pi-1)(1+b^2)}}\right). \tag{37}$$

Figure 8 shows the relationship between AUCs and the parameter b prescribed by Eqs. (5), (9), (36), and (37) for

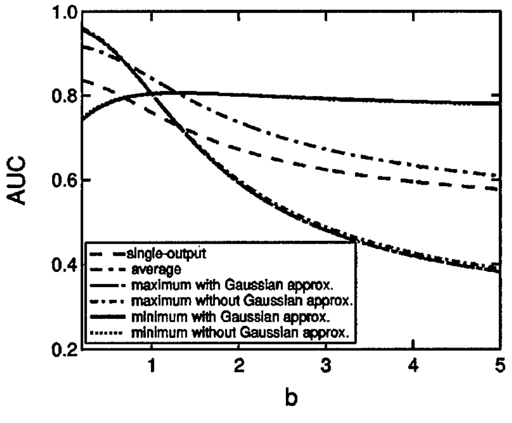

FIG. 8. AUC versus b for the *single-output*, the *average*, the *maximum* and the *minimum* derived from two independent outputs ($n=2$) that arise from the binormal distribution pair $N(0,1)$ and $N(a/b,1/b)$ with $a=1$, with and without the Gaussian approximation to the results of the *maximum* and the *minimum*.

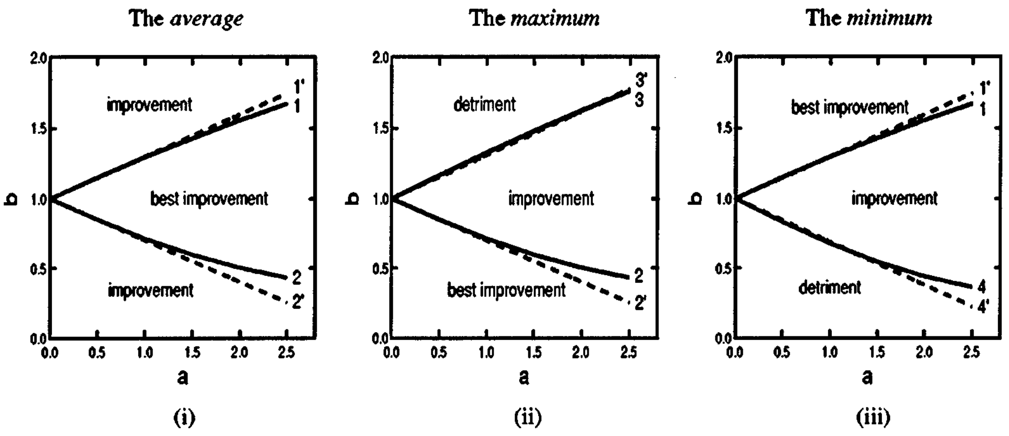

FIG. 9. Comparison of the same curves as shown in Fig. 5 but calculated with Gaussian approximation of the results from the *maximum* and the *minimum*. Curves 1, 2, 3 and 4 (solid lines) were calculated without the Gaussian approximation (same as in Fig. 5). Curves 1′, 2′, 3′ and 4′ (dashed lines) were calculated with the Gaussian approximation.

$a=1$ and $n=2$, along with the AUCs calculated numerically without the Gaussian approximation. One can see that the curves calculated with and without Gaussian approximation agree extremely well.

One can show analytically from Eqs. (5), (9), (36), and (37) that the intersection points for curves calculated with the Gaussian approximation in Fig. 8 are given by

$$b=1 \quad \text{for the } \textit{maximum} \text{ and the } \textit{minimum}, \tag{38}$$

$$(\sqrt{2(\pi-1)}-\sqrt{\pi})a+b-1=0$$
$$\text{for the } \textit{maximum} \text{ and the } \textit{average}, \tag{39}$$

$$(\sqrt{2(\pi-1)}-\sqrt{\pi})a-b+1=0$$
$$\text{for the } \textit{minimum} \text{ and the } \textit{average}, \tag{40}$$

$$(\sqrt{\pi-1}-\sqrt{\pi})a+b-1=0$$
$$\text{for the } \textit{maximum} \text{ and the } \textit{single-output}, \tag{41}$$

$$(\sqrt{\pi-1}-\sqrt{\pi})a-b+1=0$$
$$\text{for the } \textit{minimum} \text{ and the } \textit{single-output}. \tag{42}$$

Figure 9 shows the last four of these relationships as dashed lines along with the same curves calculated numerically without the Gaussian approximations shown as solid lines (also shown in Fig. 5). The two relationships obtained with and without the Gaussian approximation agree except in the region where $a>1.6$ and $b<0.55$. This region has a *single-output* AUC value greater than 0.92, and therefore relatively little improvement can be expected from any method of combining multiple images because of the already high *single-output* AUC value. Also the AUC value changes slowly with a and b in this region; therefore, a large difference in (a,b) corresponds a small difference in the AUC value.

It is well known that the binormal model can be applied to a wide variety of practical situations through a generally unknown monotonic transformation of the data.[8,9] However, the highly linear appearance of the normal-deviate ROC curves of the *maximum* and the *minimum*, together with the excellent Gaussian approximation to their probability functions, are surprising. These observations provide additional strong evidence for the robustness of the binormal model.

C. Limitations and future work

Our analysis of multiple computer outputs per case is motivated by work in CAD, where typically multiple feature values are extracted for each image and then analyzed by a classifier. For example, Jiang *et al.*[4] used eight image features to classify microcalcifications on mammograms. When multiple images from the same patient are analyzed at the same time, it is easier to combine the classifier outputs than to treat a feature from multiple images (e.g., lesion area from MLO and CC view mammograms) as multiple features, because the latter approach would require a larger classifier. Also, combining classifier outputs allows each image to be used separately in training and testing of the classifier, thus enlarging the database to some extent. These considerations motivated the present work.

In this work, we assumed that the multiple outputs arise from the same binormal model. It is reasonable to assume the individual outputs to follow the binormal model because the binormal model has been shown to be robust across many applications.[9] However, this assumption limits the theory to applications where each individual output produces the same ROC curve. We defer the analysis of multiple outputs that do not follow the same binormal distribution pair to future work. We also assume here that the multiple outputs are statistically independent, which is not likely to hold in real applications because multiple images of the same patient are almost always correlated. Extending our present theory to correlated multiple outputs is not straightforward and therefore also will be deferred to future work. However, the results that we have obtained with these simplifications provide a mathematical understanding of the effects of the *average*, the *median*, the *maximum*, and the *minimum* on distribution width and separation, and, therefore, on the corresponding ROC curves. We expect the trend predicted by our theory to hold at least qualitatively in real applications, namely, that the *average* always improves diagnostic performance when *single-output* ROC curves are similar, and that the *maximum* and the *minimum* can outperform the *average* for small b and for large b, respectively.

V. CONCLUSION

We have investigated theoretically the diagnostic performance of using the *average*, the *median*, the *maximum* or the *minimum* to combine multiple computer outputs that are obtained from multiple images of the same patient in CAD, assuming that the computer outputs follow the same binormal model and that the multiple computer outputs are statistically independent. In this situation, the *average* also follows the binormal model, whereas the *maximum* and the *minimum* were found to follow the binormal model approximately. As expected, the *average* always improves the ROC curve, because it reduces the overlap between the distributions associated with actually-negative and actually-positive cases. However, the *maximum* and the *minimum* can also improve the ROC curve, and depending on the *single-output* distributions, can outperform the *average* in certain situations. We have identified the situations in which one should use the *average*, the *maximum* or the *minimum* in order to obtain the highest ROC curve. To the extent that these results apply qualitatively also to correlated outputs, as we expect, our theory provides guidance concerning how best to combine multiple outputs to improve diagnostic performance. This guidance is appropriate for applications such as replicated readings of radiographs by a single radiologist or by a group of radiologists with similar accuracy, and for combination of CC and MLO views in mammography.

ACKNOWLEDGMENTS

This work was supported in part by the NCI/NIH through Grant No. R01-CA92361 and by the U.S. Army Medical Research and Materiel Command through Grant No. DAMD17-00-1-0197. C. E. Metz is a shareholder in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

APPENDIX A: DERIVATION OF A_z AS A FUNCTION OF a AND b

Under the binormal model, $\mathrm{TPF}(x)=\Phi(a-bx)$ and $\mathrm{FPF}(x)=\Phi(-x)$ as denoted in Eqs. (3) and (4). Therefore,

$$A_z=\int_0^1 \mathrm{TPF}d(\mathrm{FPF})$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Phi(a-bx)e^{-(1/2)x^2}dx$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-bx}e^{-(1/2)t^2}dt\right)e^{-(1/2)x^2}dx.$$

Letting $y=t-a+bx$, we have

$$A_z=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{0}e^{-(1/2)(y+a-bx)^2}dy\right)e^{-(1/2)x^2}dx$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-(1/2)(b^2+1)(x-b(y+a)/(b^2+1))^2}dx$$

$$\times\int_{-\infty}^{0}e^{-(1/2)(y+a)^2/(1+b^2)}dy$$

$$=\frac{1}{\sqrt{2\pi}\sqrt{1+b^2}}\int_{-\infty}^{0}e^{-(1/2)(y+a)^2/(1+b^2)}dy$$

$$=\Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{A1}$$

APPENDIX B: THE MEANS AND STANDARD DEVIATIONS OF THE MAXIMUM AND THE MINIMUM OF TWO INDEPENDENT IDENTICALLY DISTRIBUTED NORMAL RANDOM VARIABLES

For two random variables x_1 and x_2 that are drawn independently from $N(\mu,\sigma)$, x_1+x_2 and x_1-x_2 are independent.[14] We also have[15]

$$x_{\max}=\max\{x_1,x_2\}=\tfrac{1}{2}(x_1+x_2)+\tfrac{1}{2}|x_1-x_2|, \tag{B1}$$

and

$$x_{\min}=\min\{x_1,x_2\}=\tfrac{1}{2}(x_1+x_2)-\tfrac{1}{2}|x_1-x_2|. \tag{B2}$$

Let $y=x_1-x_2$, then y is a random variable of normal distribution $N(0,\sqrt{2}\sigma)$. Therefore, the mean of $|y|$ is

$$\overline{|y|}=\frac{1}{\sqrt{2\pi}\sqrt{2}\sigma}\int_{-\infty}^{\infty}|y|e^{-(1/2)(y/\sqrt{2}\sigma)^2}dy$$

$$=\frac{1}{\sqrt{2\pi}\sqrt{2}\sigma}\left(\int_{-\infty}^{0}(-y)e^{-(1/2)(y/\sqrt{2}\sigma)^2}dy\right.$$

$$\left.+\int_{0}^{\infty}ye^{-(1/2)(y/\sqrt{2}\sigma)^2}dy\right)$$

$$=\frac{2}{\sqrt{2\pi}\sqrt{2}\sigma}\int_{0}^{\infty}ye^{-(1/2)(y/\sqrt{2}\sigma)^2}dy=\frac{2\sigma}{\sqrt{\pi}} \tag{B3}$$

and the variation of $|y|$ is

$$\mathrm{Var}(|y|)=\overline{|y|^2}-(\overline{|y|})^2$$

$$=\overline{y^2}-(\overline{|y|})^2=2\sigma^2-\left(\frac{2\sigma}{\sqrt{\pi}}\right)^2=2\sigma^2-\frac{4\sigma^2}{\pi} \tag{B4}$$

so the mean of $x_{\max}$ is

$$\overline{x_{\max}}=\tfrac{1}{2}(\overline{x_1}+\overline{x_2})+\tfrac{1}{2}\overline{|y|}=\tfrac{1}{2}(\mu+\mu)+\frac{\sigma}{\sqrt{\pi}}=\mu+\frac{\sigma}{\sqrt{\pi}}. \tag{B5}$$

Since x_1+x_2 and x_1-x_2 are independent, the standard deviation of $x_{\max}$ is

$$\begin{aligned}\mathrm{SD}_{x_{\max}} &= \sqrt{\tfrac{1}{4}\mathrm{Var}(x_1+x_2)+\tfrac{1}{4}\mathrm{Var}(|y|)}\\ &= \sqrt{\frac{1}{4}(\sigma^2+\sigma^2)+\frac{1}{4}\left(2\sigma^2-\frac{4\sigma^2}{\pi}\right)}=\sqrt{1-\frac{1}{\pi}}\sigma.\end{aligned} \tag{B6}$$

Similarly, the mean and standard deviation of $x_{\min}$ are given by

$$\overline{x_{\min}}=\mu-\frac{\sigma}{\sqrt{\pi}}, \tag{B7}$$

$$\mathrm{SD}_{x_{\min}}=\sqrt{\left(1-\frac{1}{\pi}\right)}\sigma. \tag{B8}$$

[1] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).

[2] C. E. Metz and J. Shen, "Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis," Med. Decis Making **12**, 60–75 (1992).

[3] R. G. Swensson, J. L. King, W. F. Good, and D. Gur, "Observer variation and the performance accuracy gained by averaging ratings of abnormality," Med. Phys. **27**, 1920–1933 (2000).

[4] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vybomy, and K. Doi, "Malignant and benign clustered microcalcifications: Automated feature analysis and classification," Radiology **198**, 671–678 (1996).

[5] H. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," Med. Phys. **25**, 2007–2019 (1998).

[6] Z. Huo, M. L. Giger, and C. J. Vybomy, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," IEEE Trans. Med. Imaging **20**, 1285–1292 (2001).

[7] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Invest. Radiol. **14**, 109–121 (1978).

[8] J. A. Swets, "Form of empirical ROCs in discrimination and diagnostic tasks: Implication for theory and measurement of performance," Psychol. Bull. **99**, 181–198 (1986).

[9] J. A. Hanley, "The robustness of the binormal assumptions used in fitting ROC curves," Med. Decis Making **8**, 197–203 (1988).

[10] C. E. Metz, "Statistical analysis of ROC data in evaluation of diagnostic performance," in *Multiple Regression Analysis: Applications in Health Sciences*, edited by D. Herbert and R. Myers (American Institute of Physics, New York, 1986), pp. 365–384.

[11] C. E. Metz, B. A. Herman, and J. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continously-distributed data," Stat. Med. **17**, 1033–1053 (1998).

[12] A. Papoulis, *Probability, Random Variable, and Stochastic Process* (McGraw–Hill, New York, 1991).

[13] J. T. McClave and T. Sincich, *Statistics* (Prentice–Hall, Englewood Cliffs, 2000).

[14] J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury, New York, 1995).

[15] H. A. David and H. N. Nagaraja, *Order Statistics* (Wiley, New York, 2003).

Radial gradient-based segmentation of mammographic microcalcifications: Observer evaluation and effect on CAD performance

Sophie Paquerault,[a)] Laura M. Yarusso, John Papaioannou, Yulei Jiang,
and Robert M. Nishikawa
Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC 2026, Chicago, Illinois 60637

(Received 22 September 2003; revised 7 May 2004; accepted for publication 12 May 2004; published 27 August 2004)

Precise segmentation of microcalcifications is essential in the development of accurate mammographic computer-aided diagnosis (CAD) schemes. We have designed a radial gradient-based segmentation method for microcalcifications, and compared it to both the region-growing segmentation method currently used in our CAD scheme and to the watershed segmentation method. Two observer studies were conducted to subjectively evaluate the proposed segmentation method. The first study (A) required observers to rate the segmentation accuracy on a 100-point scale. The second observer evaluation (B) was a preference study in which observers selected their preferred method from three displayed segmentation methods. In study A, the observers gave an average accuracy rating of 88 for the radial gradient-based and 50 for the region-growing segmentation method. In study B, the two observers selected the proposed method 56% and 62% of the time. We also investigated the effect of the proposed segmentation method on the performance of computerized classification scheme in differentiating malignant from benign clustered microcalcifications. The performances of the classification scheme using a linear discriminant analysis (LDA) or a Bayesian artificial neural network classifier both showed statistically significant improvements when using the proposed segmentation method. The areas under the receiver-operating characteristic curves for case-based performance when using the LDA classifier were 0.86 with the proposed segmentation method, 0.80 with the region-growing method, and 0.83 with the watershed method. © *2004 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1767692]

Key words: computer-aided diagnosis, mammography, calcifications, segmentation, classification, observer evaluation

I. INTRODUCTION

Screening mammography is the most effective technique for detecting breast cancer in its early stages.[1] However, the interpretation of mammograms for breast cancer diagnosis is a difficult task. Currently, 10%–30% of breast cancers are missed at mammography screening either due to technical error, lesion subtlety, misinterpretation, or oversight.[2,3] A significant number of false positive biopsies are also performed. These are costly and cause unnecessary trauma to the patient. Although various techniques are being developed to improve the sensitivity and specificity of breast cancer detection and characterization, mammography is the only technique capable of detecting and characterizing breast cancer that is developing as clustered microcalcifications.[4] Computer-aided diagnosis (CAD) schemes that automatically detect and characterize abnormal lesions on mammograms have been designed to provide a second opinion to radiologists. It has been demonstrated that reading with CAD can significantly reduce the error rate under simulated mammography screening conditions[5,6] and lower the number of biopsy recommendations for benign breast lesions.[7–9] However, the accuracy of CAD schemes can be improved, allowing radiologists to further improve their management decisions.

CAD schemes for detection or characterization of mammographic lesions generally consist of three stages. As with many pattern recognition problems, the first stage of CAD schemes is the identification and segmentation of suspicious breast lesions. The second stage is feature extraction, in which useful information about the segmented lesions or objects is extracted. The third stage is a classifier that weights relevant features in order to differentiate true positive from false positive detections, or malignant from benign lesions. If any of these three stages is individually improved, it is likely to increase the overall accuracy of the CAD scheme.

Many approaches to further improve the accuracy of CAD schemes for the detection and classification of mammographic lesions have been developed. Most of these approaches use extraction of morphological, textural, geometrical features from the segmented lesions and are based on concepts of computer vision and pattern recognition.[10] Alternate approaches attempt to incorporate additional features, including information on the mammographic parenchymal patterns and *a priori* information on the patient demographic information, medical, and family history.[11] Also, the addition of the Breast Imaging Reporting and Data System (BI-RADS) descriptors[12] provided by radiologist has demonstrated an improvement of the computerized classification scheme accuracy when combined with the computer-extracted lesion features.[13] Although these approaches can potentially be used to improve the accuracy of CAD schemes, there is further need for improvements to the accuracy and reproducibility of CAD schemes.

Veldkamp and Karssemeijer[14] demonstrated that microcalcification segmentation strongly influences the accuracy of the shape and size features and thus the results of classifying clustered microcalcifications into malignant or benign categories. In the past ten years, many segmentation techniques developed in diverse imaging and signal processing fields have been applied to identify the contours of mammographic microcalcifications. Four popular approaches for segmentation are (1) thresholding methods, (2) edge-based methods, (3) region-based methods, and (4) connectivity-preserving relaxation methods. Thresholding and region-growing techniques are the most common segmentation methods used in CAD schemes. Chan *et al.*[15] used a local noise-based threshold for region-growing in a signal-enhanced image. Jiang *et al.*[16] applied a background trend correction and used a signal-dependent threshold for region growing. Other approaches have been applied to mammographic microcalcifications and to objects in phantom images.[17–19]

In this study, we investigated a novel approach for automated microcalcification segmentation using a radial gradient-based method. The intention was to develop a segmentation method that is independent of the imaging system, and to segment lesions accurately in both standard, compression, and magnification views. Hopefully, parameter adjustments between modalities and view types would not be required. We performed observer evaluations of the accuracy of the proposed method and compared it to both the region-growing segmentation method currently used in the CAD scheme developed at the University of Chicago and to the watershed method. We also incorporated the proposed segmentation method into a computerized classification scheme and studied its effect on the accuracy of computer performance in differentiating malignant from benign clustered microcalcifications. To date, we are unaware of any studies which evaluate accuracy of microcalcification segmentation with respect to observers' opinions. To the best of our knowledge, there are also no published reports on the effects of segmentation on a CAD classification scheme.

II. MATERIALS AND METHODS

We have developed a radial gradient-based segmentation method to accurately identify the contours of individual microcalcifications. The accuracy of the proposed method was evaluated through observer studies and compared to two other segmentation methods. The proposed segmentation method was also incorporated into a computerized scheme for classification of malignant versus benign clustered microcalcifications in order to determine if the computer performance improved with the proposed segmentation method.

A. Database

The database for this study consisted of 144 microcalcification clusters from mammograms of 76 patients. The cases were selected from patient files in the Department of Radiology at the University of Chicago. Each case was biopsy proven. Fifty-nine mammograms contained malignant microcalcification clusters, 85 were benign. The mammograms were digitized with a (LUMISYS) laser scanner at a pixel size of 100 μm and 12-bit gray scale, which is consistent with pixel size and bit depth used in the current CAD scheme. The histogram of the number of manually identified microcalcifications per cluster is shown in Fig. 1.

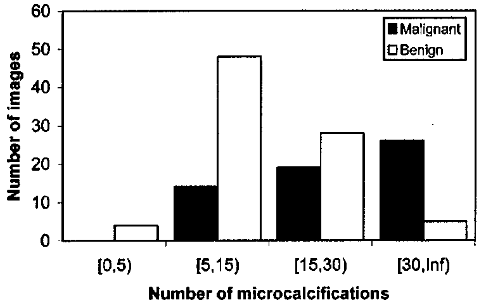

FIG. 1. Histogram of the number of microcalcifications per cluster on our database of 144 mammographic images.

A subset of the database described above was used for the observer evaluation of the proposed radial gradient-based segmentation method because of the length of time required for the observer study. The subset was chosen to be representative of the entire database and contained microcalcifications with various sizes and shapes. This subset consisted of 50 microcalcification clusters from mammograms of 31 patients.

B. Segmentation methods

The three different segmentation methods we evaluated were the proposed radial gradient-based segmentation method, the region-growing method used in the current CAD scheme, and the watershed method. All segmentation methods begin with manually identified seed points which were carefully selected to represent the centers of the individual microcalcifications. This manual selection could be replaced by an automated detection scheme to identify microcalcification locations. However, in this study, we chose to use manually identified seed points to avoid incorporating false-positive computer detections into analysis by the CAD classification scheme, and also to ensure that every microcalcification was analyzed.

1. Region-growing segmentation method currently used in CAD scheme

The segmentation method used in the current CAD scheme is a region-growing method, which is among the most popular methods used.[20] Preprocessing is performed on the image before microcalcification segmentation. This preprocessing stage consists of an application of a background trend correction using a third-degree polynomial fit in the 1×1 cm^2 region centered on the manually identified seed point. A four-point connectivity region-growing technique based on two consecutive gray level thresholds is then used to identify member pixels that belong to each microcalcification. The thresholds are defined based on the local gray level characteristics. The first threshold is 50% of the maximum intensity in a 5×5 pixel region around the seed point; the second threshold is 50% of this same maximum intensity minus the residual background offset, which is computed as the mean intensity in the $1\times1\ \mathrm{cm}^2$ region excluding pixels retained after the first threshold. Thus, starting at a manually identified seed point (x_s, y_s) with preprocessed image intensity values of $f(x_s, y_s)$, all four-connected neighboring pixels are visited and pixels are added if their intensities $f(x,y)$ are above the defined threshold. For both thresholds, this process is repeated until all four-connected pixels with intensity values above the threshold are found. The microcalcification contour is completely identified when no more neighboring pixels can be added or when the segmented microcalcifcation size reaches a maximum area of 121 pixels at 100 μm resolution.

2. Watershed segmentation method

The watershed method is well known in topography.[21] It is a powerful mathematical morphology tool for segmentation that does not involve *a priori* information. It treats images as topographic surfaces. In our case, mammographic microcalcifications correspond to small elevations. Considering the gradient magnitude at each pixel as the surface height, regions are formed by simulating a flooding of the image that starts at a local minimum of the gradient image function. The contour of each segmented region stops advancing when neighboring flooding regions meet. The principal advantage of the watershed segmentation method over other methods is that it requires no parameter input, including thresholds.[22] In this study, the watershed method was applied in a manner such that neither preprocessing nor background trend correction was necessary. For each microcalcification, the contour was identified by applying the flooding simulation to an inverse gray-level image starting at the manually identified seed point.

3. Radial gradient-based segmentation method

We developed and implemented a novel gradient-based segmentation method. It does not require any preprocessing of the image or background trend correction. It is independent of the pixel size of the imaging system and requires no input parameters such as thresholds. The radial gradient-based method individually segments microcalcifications starting from each manually identified seed point. For each seed point, the mammogram is transformed into polar coordinates. Each radial distance and angle from the seed point under consideration corresponds to a bilinear-interpolated gray level in the image. This transformation is required to produce continuous microcalcification contours in contrast to the discrete contours output by the two segmentation methods previously described. From this transformed image, the radial gradient map is computed. For each pixel with intensity $f(r,\theta)$, the radial gradient $g(r,\theta)$ is expressed as

$$g(r,\theta)=\frac{f(r,\theta)-f(0,\theta)}{r}, \tag{1}$$

where r is the radial distance of a given pixel to the seed point, and θ is the angle from the horizontal image axis. The quantity $f(0,\theta)$ is the gray level intensity at the seed point and is constant for all angles θ.

A pixel is considered to be a candidate member of the microcalcification contour if it provides the minimal radial gradient G, as described below. To avoid local minima, a restriction is imposed so that all microcalcification contour pixels are members of a unique and continuous minimal radial gradient contour, also called the minimal radial gradient road map. Thus, we search the candidate set L of consecutive microcalcification contour pixels that minimizes

$$G=\sum_{i\in L} g(r_i,\theta_i), \quad L=\{(r_i,\theta_i), \quad i=1,\ldots,N\}, \tag{2}$$

where N is the number of points sampled on the candidate contour and set to 180.

We first extract the optimal candidate set L of contour pixels from which the global radial gradient minima G surrounding the seed point is obtained. This optimal set L imposes an upper limit on the distance of a contour pixel from the seed point. It is possible that this optimal set of points might be discontinuous and many combinations of pixel contours consistent with this limit must be tested. Therefore, to avoid a computational burden in the search, we limit the road map search to the set M, a sample of nine points equally distributed between 0 and 2π, given by

$$M=\{(r_j,\theta_j), \quad j=1,\ldots,9\}. \tag{3}$$

For each possible set M of candidate contour pixels, the path of the road map defined by the nine points is determined using a first-order polynomial fit between the points. Therefore, for each possible set M, a continuous partition L is defined as a potential road map of the microcalcification contour. For each new partition, the radial gradient contour function G is calculated.

Consider G_o to be the initial value of the radial gradient road map function G before the first iteration and M to be the new set of nine points identified from the image. If the newly calculated minimal radial gradient G is lower than G_o, then the new set M of nine points is retained as potential contour pixels and G_o is updated by the new G value. If the newly calculated G is greater than or equal to G_o, then M is rejected. The iterative process stops when no more changes in the partition M can further minimize G. At this point, the continuous and closed contour has been completely determined. To avoid drawing too many possible partitions M, we incorporate information about the spatial extent of previously segmented microcalcifications.

Other complex methods could be attempted using active contour or probabilistic techniques.[23] However, parameters required in these methods are difficult to define due to the small size of the microcalcifications. In addition, adjustments of parameters in these methods may be complicated by dependence on the imaging system used, especially its corresponding spatial resolution.

C. Observer evaluation of segmentation accuracy

The accuracies of the three segmentation methods were qualitatively evaluated in observer studies. Since it is extremely difficult and very time consuming for radiologists to manually outline the contour of each microcalcification in a cluster, we instead conducted two different observer studies using two separate user interfaces. These studies required visual assessment only and no diagnostic or management decision, therefore, the participation of radiologists was not necessary. Two observers participated in each study. The participants were one radiologist and three scientists. The scientists were experienced in mammography and medical imaging with 5–10 years of work in the field. No observers were involved in the development of any of the segmentation methods. They were not told about the difference in the segmentation methods. Observer bias was further avoided by not visually identifying any of the three methods for observers either before or during the study.

In both user interfaces, the digitized mammograms were displayed with the manually identified center locations of the microcalcifications superimposed on the images. This was done so that observers could mentally visualize the contours that they would draw for each microcalcification in the cluster and to ensure that all observers analyzed the same set of microcalcifications. The observers were then shown the computer-extracted contour superimposed on the digitized mammograms. Observers were allowed to adjust the image brightness and contrast, zoom to specific regions of the image, and highlight segmented microcalcification contours to see the degree of overlap, if any, between neighboring microcalcifications.

1. Observer study A

In observer study A, two observers rated the accuracy of segmentation for both the proposed radial gradient-based method and the region-growing method. These two segmentations were displayed by superimposing the two contours on adjacent copies of the original image. Observers were instructed to rate the accuracy of segmentation for each method based on how well the computer-segmented contour agreed with their mentally visualized contours in terms of criteria such as segmented shape and area. Observers provided a single accuracy rating for each segmentation method, which they were instructed was the average value over all microcalcifications in the cluster. The accuracy rating was on a 100-point scale. A rating of 100 implied that they observed perfect segmentation by the computer, and 0 implied completely inaccurate segmentation. These accuracy ratings were entered directly into the computer by each observer. Intraobserver consistencies were evaluated by having observers randomly review cases more than one time.

TABLE I. Eight computer-extracted features for classification of malignant and benign microcalcification clusters.

Feature index	Feature description
1	Cluster circularity
2	Cluster area
3	Number of microcalcifications
4	Average effective volume of microcalcifications
5	Relative standard deviation in effective thickness
6	Relative standard deviation in effective volume
7	Average area of microcalcifications
8	Second highest microcalcification-shape-irregularity measure in a cluster

2. Observer study B

Observer study B was a preference study between the three segmentation methods. Observers were instructed to select their preferred method from three simultaneously displayed segmentation methods: the proposed radial gradient-based segmentation method, the watershed method, and the region-growing method. The user interface designed for this observer study simultaneously displayed three images with the three different segmentation results superimposed on the original digitized mammogram. The three displayed segmentation methods were randomly reordered from one case to the next. Observers were instructed to choose their preferred segmentation method using criteria such as the segmented shape and area. For each set of three segmentation results, the observer entered the preferred method into the computer. Intraobserver consistencies were evaluated by having observers randomly review cases multiple times.

D. Effect of segmentation method on CAD performance

We evaluated the computerized classification scheme performance when using each of the three segmentation methods. In the classification scheme for malignant versus benign clustered microcalcifications, shape and size features are extracted both for individually segmented calcifications and for the cluster. We used the same eight features reported in Jiang *et al.*[16] as detailed in Table I. The performances of these extracted features closely depend on the accuracy of the microcalcification segmentation used, and therefore, we expected that using more precise segmentation would improve the accuracy of the CAD classification scheme. To determine the potential improvement to the CAD scheme, we compared classification performance based on the proposed radial gradient-based method, the watershed method, and the region-growing technique currently used in the classification scheme.

A linear discriminant analysis (LDA) classifier and a Bayesian artificial neural network (BANN) classifier were employed separately in this study. We chose to evaluate the CAD classification performance using both LDA and BANN classifiers in order to draw more robust conclusions on the CAD scheme performance when using each of the three segmentation methods. The LDA classifier has the advantage of

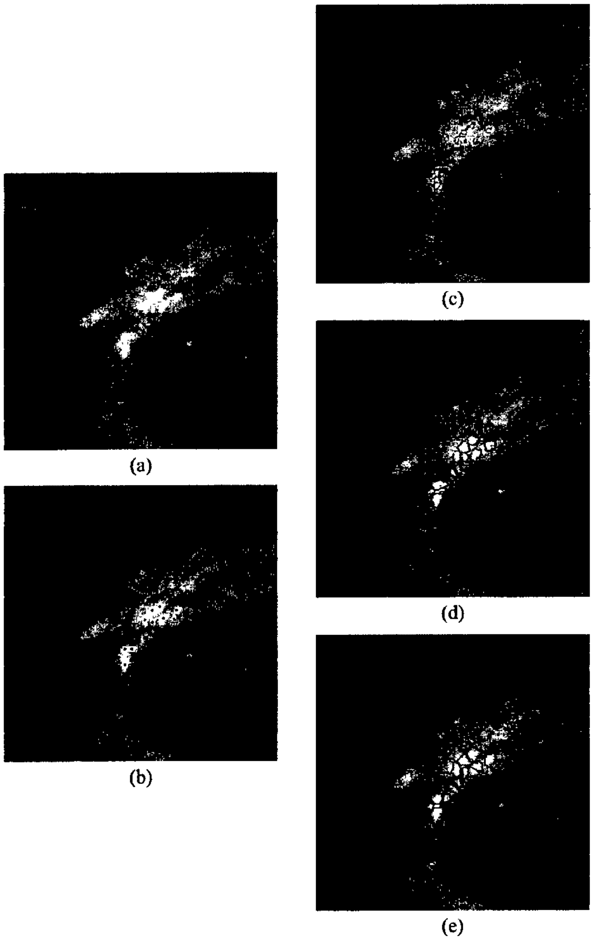

FIG. 2. Example of segmentation results for a malignant microcalcification cluster: (a) original mammogram, (b) superimposed manually identified microcalcification locations, (c) region-growing method, (d) radial gradient-based method, and (e) watershed method contours.

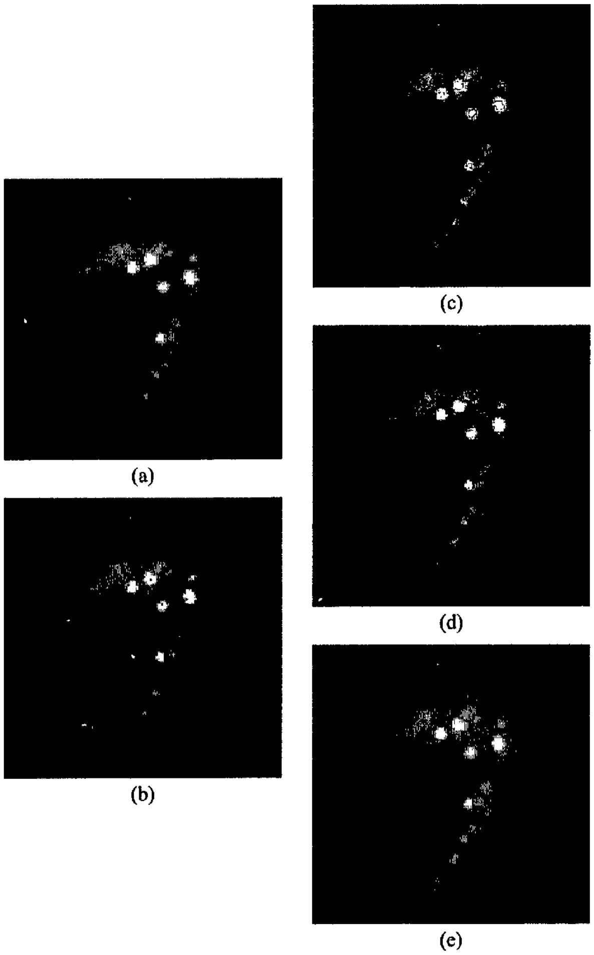

FIG. 3. Example of segmentation results for a benign microcalcification cluster: (a) original mammogram, (b) superimposed manually identified microcalcification locations, (c) region-growing method, (d) radial gradient-based method, and (e) watershed method contours.

computational speed in the training stage, and is often a preferred choice when the number of available training samples is small.[24] The BANN classifier avoids the overtraining problem and also incorporates uncertainty about classifier coefficients into its output.[25] The LDA and BANN classifiers were implemented using the eight extracted features as inputs. The leave-one-out technique was used to train these classifiers. Both classifiers output the likelihood of malignancy for each cluster. The accuracies of the LDA and BANN classifiers were evaluated separately with receiver-operating characteristic (ROC) curve methodology and the area under the ROC curve (A_z) was used as a performance index.

III. RESULTS

A. Segmentation methods

The three segmentation methods were applied to the microcalcifications in our database. Figures 2(a) and 3(a) show malignant and benign mammographic microcalcification clusters, respectively. The corresponding manually identified locations of the microcalcifications are shown in Figs. 2(b) and 3(b). The segmentation from the region-growing method used in the current CAD scheme is illustrated in Figs. 2(c) and 3(c). The proposed radial gradient-based segmentation method is illustrated in Figs. 2(d) and 3(d), and the watershed segmentation method in Figs. 2(e) and 3(e). Three main differences between the resulting contours from the three segmentation methods are evident in Figs. 2 and 3. The sizes of the segmented calcifications typically were smaller for the region-growing method, larger when using the watershed method, and intermediate for the proposed radial gradient-based method. The shapes of the microcalcification contours were also very different depending on the segmentation method used. The proposed radial gradient-based method tended to extract smoother, rounder contours than the somewhat more irregular and jagged contours extracted when using the other two segmentation methods. The overlaps between segmented microcalcifications were also markedly different between the three methods. This is because both the watershed and proposed radial gradient-based methods in-

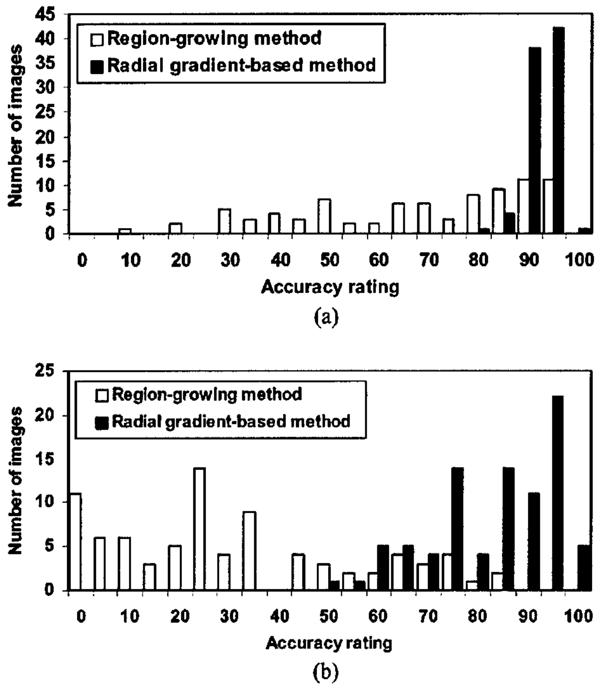

FIG. 4. Histograms of accuracy ratings of (a) observer 1 and (b) observer 2.

clude criteria that eliminate overlap between neighboring microcalcifications, while the region-growing method allows for such overlap.

B. Observer evaluation

Observer study A required observers to provide accuracy ratings on a 100-point rating scale for both the proposed radial gradient-based segmentation method and the region-growing method currently used in CAD scheme. Figure 4 shows histograms of the observers' accuracy ratings after reviewing all images. The accuracy ratings for the region-growing method were 67.9±22.9 for the first observer and 31.4±24.5 for the second observer. For the radial gradient-based method, the accuracy ratings were 91.7±3.2 for the first observer and 83.2± 12.4 for the second observer. Cumulative histograms of observers' accuracy ratings when using the accuracy rating as an operating threshold are shown in Fig. 5; a microcalcification cluster is considered to be adequately segmented if the observer rating is above or equal to the operating threshold. For a threshold accuracy rating of 50, both observers preferred the proposed radial gradient-based segmentation method to the region-growing method. At the same threshold accuracy rating, the proposed radial gradient-based method was considered by both observers to adequately segment the entire database. At the same threshold, the two observers considered the region-growing method to adequately segment 77 and 25 % of the cases, respectively.

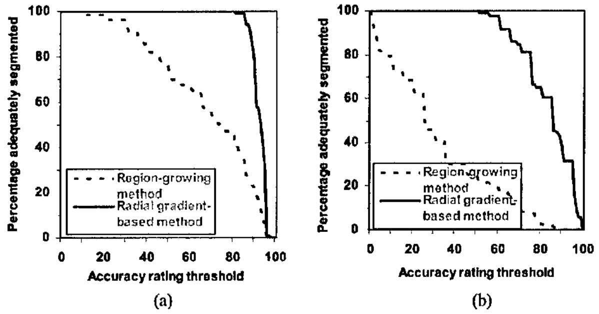

FIG. 5. Cumulative histograms of accuracy ratings of (a) observer 1 and (b) observer 2, when using the accuracy rating as an operating threshold. Images with accuracy ratings above or equal to the operating threshold are considered as being adequately segmented.

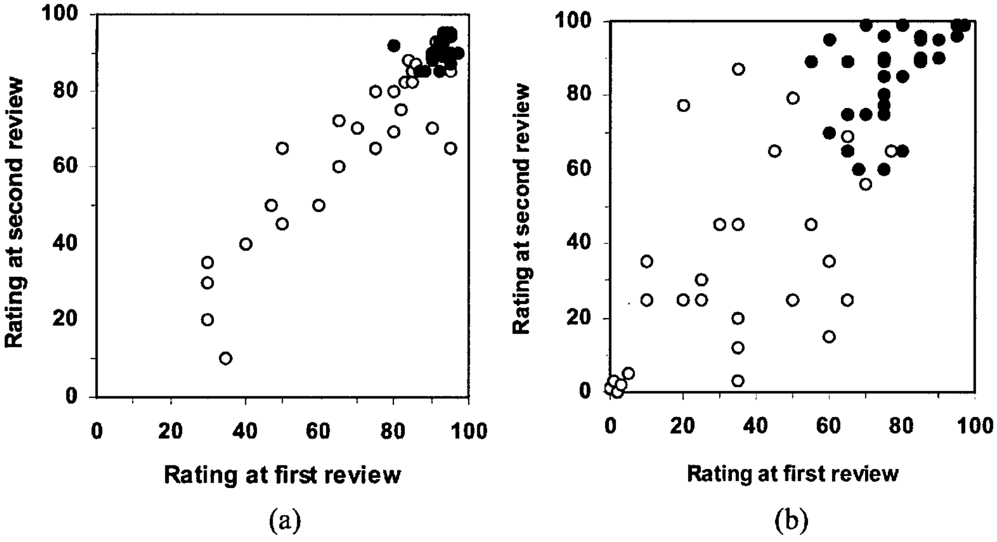

FIG. 6. Intraobserver variability in rating the accuracies of the region-growing method and the proposed radial gradient-based method: (a) observer 1 and (b) observer 2.

TABLE II. For observer study B, the preference of the two observers among the three segmentation methods. Results are displayed in terms of both numbers of cases and percentages of the database. Fifty microcalcification clusters were randomly reviewed multiple times.

	Observer 1 (all reviews)	Observer 2 (all reviews)
Region-growing	34 (20%)	31 (18%)
Radial gradient-based	94 (56%)	104 (62%)
Watershed	41 (24%)	34 (20%)

Intraobserver variability in rating the two segmentation techniques was evaluated and is illustrated in Fig. 6. The correlation between the accuracy ratings given at first and second reviews was computed. For the region-growing method, the correlation was 0.93 for the first observer and 0.64 for the second observer. For the proposed radial gradient-based method, the correlations were 0.97 for the first observer and 0.81 for the second observer. Interobserver variability was also analyzed. The correlation between the two observers' accuracy ratings was 0.77 when using accuracy ratings given at first review and 0.76 for all accuracy ratings. Student's t tests for paired data were performed and demonstrated a statistically significant difference in accuracy ratings between the region-growing and proposed segmentation methods.[26] For accuracy ratings given at first review, the t value was 3.4 at the 0.001 level of significance, while t estimated was 8.53. For all accuracy ratings, the t value was 3.29 at the 0.001 level of significance, while t estimated was 15.27.

Observer study B consisted of selecting the preferred segmentation method from the results displayed for the watershed, the region-growing, and the radial gradient-based methods. When reviewing the image for the first time, the two observers preferred the proposed radial gradient-based method 84 and 90 % of the time. The results of observers' selection for all images reviewed are shown in Table II. When accounting for the display of the same cases multiple times, the proposed radial gradient-based method was preferred by the observers 56 and 62 % of the time. Their other selections were equally distributed between the other two segmentation methods. The percent agreements between the two observers' preferences when accounting for reviewing the same cases multiple times are shown in Table III. The maximum percent agreement occurred with the radial gradient-based segmentation method. It was 78% when reviewing the image for the first time and 41% when accounting for all image reviews. Their other percent agreements and disagreements were equally distributed between the other possible matches of segmentation method preferences.

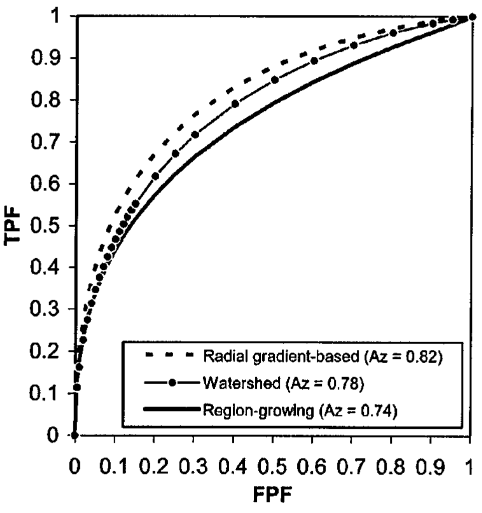

FIG. 7. Image-based performances for classification of malignant vs benign microcalcification clusters using a LDA classifier to merge the eight features extracted from segmented microcalcifications.

C. Effect of segmentation method on CAD performance

The ROC curves for CAD classification scheme when using each of the three different segmentation methods with the LDA classifier are shown in Fig. 7 for image-based performance and Fig. 8 for case-based performance. Figures 9 and 10 illustrate the ROC curves for image-based and case-based performances, respectively, when using the BANN classifier. The improvement in CAD performance for the classification of malignant versus benign microcalcification clusters when using the proposed radial gradient-based segmentation method was independent of the classifier used to merge the features. For the BANN classifier, A_z for the case-based performance was 0.82 when using the proposed radial gradient-based method, 0.73 when using the region-growing method, and 0.77 when using the watershed method. For the LDA classifier, A_z for the case-based performance was 0.86 when using the proposed radial gradient-based method, 0.80 when using the region-growing method, and 0.83 when using the watershed method. The difference in classification performance between the proposed radial gradient-based method

TABLE III. For observer study B, percent agreement between the two observers' preferences when accounting for reviewing the same cases multiple times.

		Observer 1		
		Radial gradient-based	Region-growing	Watershed
Observer 2	Radial gradient-based	41	7	7
	Region-growing	11	5	5
	Watershed	9	7	8

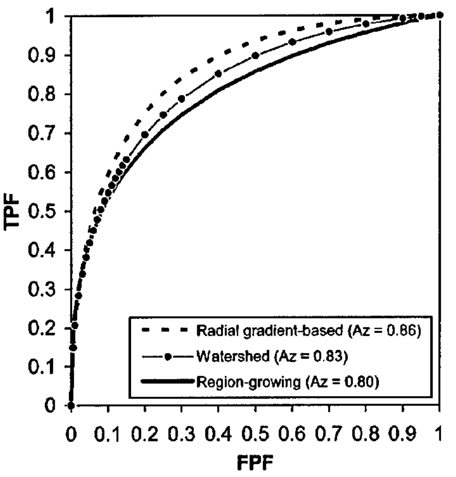

FIG. 8. Case-based performances for classification of malignant vs benign microcalcification clusters using a LDA classifier to merge the eight features extracted from segmented microcalcifications.

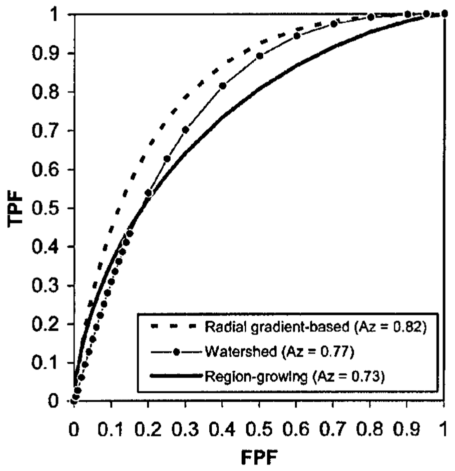

FIG. 10. Case-based performances for classification of malignant vs benign microcalcification clusters using a BANN classifier to merge the eight features extracted from segmented microcalcifications.

and the region-growing method currently used in the CAD scheme was found to be statistically significant.[27] The comparisons between other pairs of segmentation methods did not demonstrate differences that were statistically significant. The p-values are summarized in Table IV.

IV. DISCUSSION

In this research, we developed and applied a radial gradient-based segmentation method for mammographic microcalcifications, which in turn improved the computerized classification of malignant versus benign microcalcification clusters. The results of accuracy rating and preference observer studies showed that the proposed radial gradient-based method was preferred by observers. This indicates that our proposed segmentation method closely matches human visual criteria for segmentation of microcalcifications such as size, shape, and nonoverlapping criteria for segmented microcalcifications. Performance of the CAD scheme improved when incorporating the proposed segmentation method. Similar trends were observed when using either a LDA or a BANN classifier.

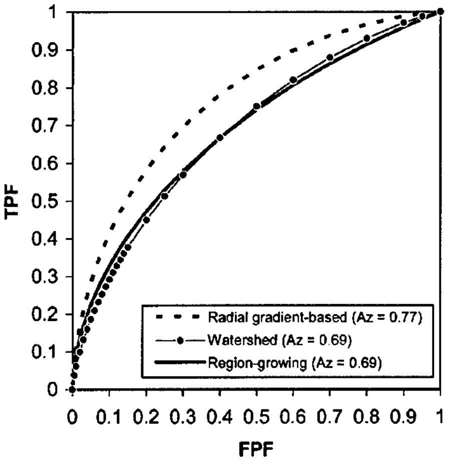

FIG. 9. Image-based performances for classification of malignant vs benign microcalcification clusters using a BANN classifier to merge the eight features extracted from segmented microcalcifications.

In this study, we chose to design a segmentation technique that does not require any preprocessing of the mammogram, background trend correction, or corrections for imaging system parameters including pixel size. Unlike the current region-growing method, the proposed radial gradient-based method has the advantage of being independent of pixel size and other parameters of the imaging system. Background correction and other preprocessing stages were not components of either our proposed method or the watershed method. We expect that using these preprocessing stages may result in more accurate segmentation, specifically in cases where microcalcifications lie on the border of dense breast tissue, or occur near the skin line. In addition, background trend correction may have eliminated the tendency of the watershed segmentation to over-segment onto normal breast tissue. It is important to note that elimination of preprocessing allows a segmentation method to be more easily applied to a wide range of mammographic image types, including both conventional and magnification views acquired on screen-film or digital systems.

Manually identified locations of microcalcifications were used as seed points for segmentation. In a fully automated CAD scheme for detection and classification of mammographic lesions, this step is eliminated.[28,29] The proposed radial gradient-based segmentation method can easily be applied to these computer-identified locations. For the purposes of this study, we did not use automated detection of the microcalcification locations because it would have complicated our comparison of the proposed segmentation method with the region-growing method currently used in the CAD classification scheme. Also, it was previously shown that inputting more false positive detected microcalcifications into the

TABLE IV. Results for testing the statistical significance of differences in A_z values, or the areas under ROC curves.

Comparison			Classifier	Image-based p-value	Case-based p-value
Region-growing	vs	Radial gradient-based	LDA	0.03	0.03
Region-growing	vs	Watershed	LDA	0.18	0.38
Radial gradient-based	vs	Watershed	LDA	0.26	0.09
Region-growing	vs	Radial gradient-based	BANN	0.03	0.06
Region-growing	vs	Watershed	BANN	0.47	0.27
Radial gradient-based	vs	Watershed	BANN	0.05	0.18

computerized classification scheme resulted in decreased classification performance, when using the region-growing method.[29] Future research is necessary to determine whether the same result occurs when incorporating the proposed radial gradient-based segmentation method into the CAD scheme.

In this work, a set of eight features based on shape and size of the cluster and of the microcalcifications was used, identical to that used in the current CAD scheme for classification of malignant versus benign microcalcifications. We chose to keep this set of features constant, as changes to other stages of the CAD scheme would have made it difficult to isolate the direct influence of the segmentation stage on the CAD scheme performance. However, many studies have focused on developing and identifying features to better discriminate malignant from benign microcalcification clusters. Evaluating the proposed segmentation method's effect on other features is an area for future investigation.

V. CONCLUSION

We developed and evaluated a radial gradient-based segmentation method to more accurately extract mammographic microcalcification contours, and thereby improve computerized classification of malignant versus benign microcalcification clusters. The advantages of our proposed segmentation method include its independence from the mammographic imaging system and its independence of external parameters associated with intensity thresholds and pixel size. Two observer studies demonstrated that the microcalcification contours resulting from our proposed method were strongly preferred to contours resulting from the region-growing method or from the watershed method. The proposed radial gradient-based segmentation method was incorporated into a CAD scheme for classification of malignant versus benign microcalcification clusters, demonstrating that more accurate microcalcification segmentation resulted in statistically significant improvements to the performance of the CAD classification scheme. More studies are underway to further improve the performance of the CAD scheme by identifying and developing features that will utilize the highly accurate contours output by the proposed segmentation method.

ACKNOWLEDGMENTS

The authors are thankful to Hiroyuki Abe, M.D., Ph.D., Alexandra V. Edwards, M.A., and Li Lan, M.S., for their participation in our study. The authors are also grateful to Charles E. Metz, Ph.D., for providing the LABROC and CLABROC programs for ROC analysis, and to Matthew A. Kupinski, Ph.D., for use of BANN programs. This work was supported in part by a grant from the NIH/NCI CA92361. R.M. Nishikawa and J. Papaioannou are shareholders of R2 Technology, Inc. (Sunnyvale, CA).

[a] Electronic mail: paquerau@uchicago.edu

[1] L. Tabar, A. Gad, L. H. Holmberg, U. Ljungquist, G. Eklund, C. J. G. Fagerberg, L. Baldetorp, O. Grontoft, B. Lundstrom, J. C. Manson, N. E. Day, and F. Pettersson, "Reduction in mortality from breast cancer after mass screening with mammography," Lancet **1**(8433), 829–832 (1985).

[2] F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353–358 (1988).

[3] D. B. Kopans, "The positive predictive value of mammography," AJR, Am. J. Roentgenol. **158**, 521–526 (1992).

[4] M. Sabel and H. Aichinger, "Recent developments in breast imaging," Phys. Med. Biol. **41**, 315–368 (1996).

[5] H. P. Chan, K. Doi, C. J. Vybomy, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[6] L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vybomy, and R. A. Castelino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554–562 (2000).

[7] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," Acad. Radiol. **6**, 22–33 (1999).

[8] H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," Radiology **212**, 817–827 (1999).

[9] D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).

[10] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," Med. Phys. **25**(10), 2007–2019 (1998).

[11] Z. Huo, M. L. Giger, D. E. Wolverton, W. Zhong, S. Cumming, and O. I. Olopade, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection," Med. Phys. **27**(1), 4–12 (2000).

[12] American College of Radiology (ACR), *Breast Imaging Reporting and*

Data System (BI-RADS), 3rd ed. (American College of Radiology, Reston, 1998).

[13] J. Y. Lo, M. K. Markey, J. A. Baker, and C. E. Floyd, Jr., "Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer," AJR, Am. J. Roentgenol. **178**, 457–463 (2002).

[14] W. J. H. Veldkamp and N. Karssemeijer, "Influence of segmentation on classification of microcalcifications in digital mammography," IEEE Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine, Proceedings of the 18th Annual International Conference of the IEEE, Vol. 3, pp. 1171–1172 (Institute of Electrical and Electronic Engineers, Inc., Amsterdam, Netherlands, 1996).

[15] H.-P. Chan, B. Sahiner, K. Leung Lam, and M. A. Helvie, "Effects of pixel size on classification of microcalcifications on digitized mammograms," Proc. SPIE Medical Imaging **2710**, 30–41 (1996).

[16] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," Radiology **198**, 671–678 (1996).

[17] J. Dengler, S. Behrens, and J. F. Desaga, "Segmentation of microcalcifications in mammograms," IEEE Trans. Med. Imaging **12**(4), 634–642 (1993).

[18] I. N. Bankman, T. Nizialek, I. Simon, O. B. Gatewood, I. N. Weinberg, and W. R. Brody, "Segmentation algorithms for detecting microcalcifications in mammograms," IEEE Trans Info. Technol. Biomed **1**(2), 141–149 (1997).

[19] W. J. H. Veldkamp and N. Karssemeijer, "Accurate segmentation and contrast measurement of microcalcifications in mammograms: a phantom study," Med. Phys. **25**(7), 1102–1110 (1998).

[20] Y. Jiang, R. M. Nishikawa, M. L. Giger, K. Doi, R. A. Schmidt, and C. J. Vyborny, "Method of extracting signal area and signal thickness of microcalcifications from digital mammograms," Proc. SPIE Medical Imaging **1778**, 28–36 (1992).

[21] J. M. Gauch and S. M. Pizer, "Multiresolution analysis of ridges and valleys in gray-scale images," IEEE Trans. Pattern Anal. Mach. Intell. **15**(6), 635–646 (1993).

[22] U. Mendonca Brago Neto, W. Siquiera Neto, and A. Flavio Dias e Silva, "Mammographic calcification detection by mathematical morphology methods," Proceedings of the 3rd International Workshop on Digital Mammography, Digital Mammography '96 (Elsevier Science, 1996), pp. 263–266.

[23] S. Timp, N. Karssemeijer, and J. Hendriks, "Comparison of three different mass segmentation methods," Proceedings of the 6th International Workshop on Digital Mammography, Digital Mammography '02 (Springer-Verlag, Berlin, Heidelberg, 2002), pp. 218–222.

[24] H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: Effects of finite sample size," Med. Phys. **26**, 2654–2668 (1999).

[25] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," IEEE Trans. Med. Imaging **20**(9), 886–899 (2001).

[26] R. F. Mould, "Testing for a significant correlation: An application of the *t* test," in *Introductory Medical Statistics*, 3rd ed. (Institute of Physics Publishing, Bristol, 1998), pp. 188–191.

[27] C. E. Metz, "Fundamental ROC analysis," in *Handbook of Medical Imaging*, edited by J. Beutel, H. Kundel, and R. L. Van Metter (SPIE Press, Bellingham, 2000), pp. 751–769.

[28] M. A. Gavrielides, J. Y. Lo, and C. E. Floyd, Jr., "Parameter optimization of a computer-aided diagnosis scheme for the segmentation of microcalcification clusters in mammograms," Med. Phys. **28**, 2403–2409 (2001).

[29] M. F. Salfity, R. M. Nishikawa, Y. Jiang, and J. Papaioannou, "Improved computerized detection of individual microcalcifications to integrate cluster detection and classification schemes," Proceedings of the 6th International Workshop on Digital Mammography, Digital Mammography '02 (Springer-Verlag, Berlin, Heidelberg, 2002), pp. 411–413.

Computer Classification of Malignant and Benign Calcifications in Full-Field Digital Mammograms

Yulei Jiang, Rich S. Rana, Robert A. Schmidt, Robert M. Nishikawa, Bei Liu,

Charlene A. Sennett, James J. Chambliss, Hiroyuki Abe

Department of Radiology, The University of Chicago

Table of Content

1. Introduction

Computer-aided diagnosis (CADx) techniques have been developed to classify breast lesions as malignant or benign in digitized screen-film mammograms. Several of these techniques have achieved promisingly high levels of performance (Getty et al. 1988; Jiang et al. 1996; Baker et al. 1996; Chan et al. 1999; Veldkamp et al. 2000; Huo et al. 2002). CADx can potentially help radiologists improve biopsy recommendations by reducing the number of biopsies performed on suspicious but benign lesions and at the same time improve sensitivity of diagnostic mammography (Jiang et al. 1999). Full-field digital mammography (FFDM) offers an opportunity to image breast lesions better (Lewin et al. 2002; Skaane and Skjennald 2004). Direct digital acquisition in FFDM makes CADx convenient and logical, eliminating the extra step of film digitization and its associated image degradation (Pisano et al. 2001).

We have previously developed a CADx technique based on digitized screen-film mammograms to classify breast calcifications as malignant or benign (Jiang et al. 1996; Jiang et al. 1999). The purpose of this study was to evaluate this CADx technique on FFDM images. In this initial study, we did not optimize performance of the computer technique for FFDM images. Therefore this study was an independent evaluation of the computer technique. The only modification made to the computer technique was in computer detection of calcifications to reduce required interactive input from radiologist and to make the computer technique more convenient to use. Therefore, in addition to evaluating the performance of the computer technique on FFDM images, this study also evaluated the computer technique as it would be used clinically.

2. Materials and Methods

2.1. Cases and Readers

This study included all diagnostic mammography cases done on a General Electric Senograph 2000D FFDM unit in our institution during 2002 and the spring of 2003. We identified 49 cases of suspicious calcifications not associated with a mass for which a biopsy was performed and definitive diagnosis was available. Of these cases, 19 contained cancers (13 of which were

DCIS) and 30 contained benign calcifications. These cases were used in this study. Four radiologists read the cases retrospectively. Three radiologists were attending radiologists specialized in mammography and one radiologist was within the first several months of a mammography fellowship training.

2.2. Computer-User Interface and Observer Study

A computer-user interface including softcopy reading of mammograms was used in this study. Two standard view mammograms were displayed one at a time. After reading a mammogram, the radiologist was asked to draw a rectangular box using a computer mouse to surround calcifications in question in that image. The radiologist was instructed to draw the box large enough to enclose all calcifications that would be appropriate to target in a hypothetical event of biopsy but no larger than necessary for enclosing only the calcifications. After reviewing both images and outlining the calcifications in both images, the radiologists was asked to enter BI-RADS assessment of 2 (benign, no biopsy), 3 (probably benign, no biopsy), 4 (suspicious, biopsy), or 5 (malignant, biopsy). Once the radiologist entered a BI-RADS assessment, the computer-estimated likelihood of malignancy was displayed numerically as a percentage for each of the two images. At this point, the radiologist could repeat reviewing either image. The radiologist could also draw a new box to ask the computer repeat its calculation, potentially modifying results of the estimated likelihood of malignancy. After reviewing all the information, the radiologist was asked to enter BI-RADS assessment again. This second BI-RADS assessment based on mammogram review and computer-estimated likelihood of malignancy could differ from the first based on mammogram review alone.

2.3. Computer Estimation of Likelihood of Malignancy

Computer analysis of the calcifications identified by the radiologist consisted of two components. The first component was computer detection of the individual calcifications and the second component was computer calculation of the likelihood of malignancy. Computer detection of individual calcifications was necessary because the locations of individual calcifications were needed for calculation of the likelihood of malignancy. Computer detection of calcifications was confined within the box drawn by the radiologist and was based on a technique described elsewhere (Salfity et al. 2003). To balance between maximizing the detection of true-positive calcifications and minimizing the detection of false-positive image noise, the computer detection was run four times; each time the computer assumed the region contained a different number of calcifications (either less than 6, between 6 and 10, between 10 and 30, or more than 30). The detection technique was designed in such a way that it would select appropriate thresholds to detect a number of calcifications that fall within the assumed range of number of calcifications. The detection results of one of the four independent runs were retained as the final detection result based on a consideration of the pattern of the detection results of all four runs. For example, if an assumption that a region contained less than 6 calcifications yielded the detection of 6 calcifications and an assumption that the region contained between 6 and 10 calcifications yielded the detection of 10 calcifications, then the true number of calcifications present was considered to be likely more than 10. However, if an assumption of between 6 and 10 yielded 10 calcifications, and an assumption of between 10 and 30 calcifications yielded 18 calcifications, but an assumption of more than 30 calcifications yielded 30 calcifications, then the detection results of 18 calcifications were retained as the final detection result. In a few cases the computer failed to determine the number of calcifications. In these cases, the radiologist was asked to provide this information.

Once individual calcifications were detected within the encompassing box indicated by the radiologist, computer calculation of the likelihood of malignancy proceeded as described in detail elsewhere (Jiang et al. 1996; Jiang et al. 1999). Briefly, the computer extracted eight image features: the size and shape of the cluster of calcifications, the number, average size, average size times contrast, uniformity in contrast, uniformity in size times contrast, and a shape-linearity measure of individual calcifications. The computer then used an artificial neural network to merge these image features into an estimate of the likelihood of malignancy. This computer classification technique was originally developed on digitized screen-film mammograms (Jiang et al. 1996; Jiang et al. 1999) and was used in this study without modification: the same image features were extracted from the digital images as from digitized images, the same feature extraction techniques were used without modification, and the same artificial neural network was used without retraining. One reason for this to be possible was that the GE digital mammograms were processed from raw images through a logarithmic transformation to make the processed image look like film. Another reason was that the GE digital mammograms had a 100-micron pixel size, the same as the digitized mammograms we used previously.

2.4. Data Analysis

ROC curves were computed from the observer study data. ROC curves were computed for the BI-RADS assessments given by each radiologist reading mammograms alone. Separate ROC curves were computed for the BI-RADS assessments given by each radiologist reading mammograms with the computer-estimated likelihood of malignancy. ROC curves were also computed for the computer-calculated likelihood of malignancy. Because each radiologist independently drew a box to indicate suspicious calcifications, for a given mammogram the computer calculated a potentially different value of the likelihood of malignancy when used by different radiologists. Therefore, four ROC curves were computed for the computer-calculated likelihood of malignancy, one for each radiologist. Group ROC curves were also computed by averaging the parameters (*a* and *b)* of maximum-likelihood fitted univariate binormal ROC curves.

3. Results and Discussion

The radiologists achieved an average A_z value (area under the ROC curve) of 0.72 reviewing the mammograms alone without the computer aid. The computer calculated likelihood of malignancy had an average A_z value of 0.79. The four ROC curves of the computer-calculated likelihood of malignancy based on each radiologist's indication of calcifications (via a box over the mammogram) were highly similar. In a previous study of digitized screen-film mammograms, this computer technique achieved an A_z value of 0.80 (Jiang et al. 1999). These results are important because they suggest that the computer calculation was able to perform consistently on FFDM images as on digitized screen-film mammograms without modification to the computer technique or retraining of the neural network classifier.

The radiologists achieved an average A_z value of 0.76 reviewing the mammograms with the aid of the computer-calculated likelihood of malignancy. While this performance was better than their unaided performance, this improvement was not statistically significant, and we did not find any other differences in A_z values to be statistically significant. Relatively small numbers of cases and readers have limited the statistical power of this study and, therefore, the results should be considered as preliminary.

4. Conclusions

This preliminary study on computer classification of malignant and benign calcifications in FFDM images indicates that radiologists can interact with the computer technique easily to indicate suspicious calcifications and for the computer to estimate the likelihood of malignancy. The computer is able to perform accurately in this independent evaluation and achieve high performance on FFDM images even though the computer technique was developed and trained independently on digitized screen-film mammograms. These promising results indicate that this CADx technique can be potentially used clinically in diagnostic mammography to aid radiologists in analyzing calcifications.

5. Acknowledgements

This work was funded in part by the NCI (R01 CA92361), the US Army Medical Research and Materiel Command (DAMD 17-00-1-0197), and the NIH (9-T35-DK062719-16). R. A. Schmidt and R. M. Nishikawa are shareholders in R2 Technology, Inc. (Sunnyvale, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by the research activities.

6. References

Getty, D. J., R. M. Pickett, C. J. D'Orsi, and J. A. Swets. 1988. Enhanced interpretation of diagnostic images. *Invest Radiol.* 23:240-252.

Jiang, Y., R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi. 1996. Malignant and benign clustered microcalcifications: automated feature analysis and classification. *Radiology.* 198:671-678.

Baker, J. A., P. J. Kornguth, J. Y. Lo, and C. E. Floyd, Jr. 1996. Artificial neural network: improving the quality of breast biopsy recommendations. *Radiology.* 198:131-135.

Chan, H. P., B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal. 1999. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology.* 212:817-827.

Veldkamp, W. J., N. Karssemeijer, J. D. Otten, and J. H. Hendriks. 2000. Automated classification of clustered microcalcifications into malignant and benign types. *Med Phys.* 27:2600-2608.

Huo, Z., M. L. Giger, C. J. Vyborny, and C. E. Metz. 2002. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology.* 224:560-568.

Jiang, Y., R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi. 1999. Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol.* 6:22-33.

Lewin, J. M., C. J. D'Orsi, R. E. Hendrick, L. J. Moss, P. K. Isaacs, A. Karellas, and G. R. Cutter. 2002. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. *AJR Am J Roentgenol.* 179:671-677.

Skaane, P., and A. Skjennald. 2004. Screen-Film Mammography versus Full-Field Digital Mammography with Soft-Copy Reading: Randomized Trial in a Population-based Screening Program--The Oslo II Study. *Radiology.* 232:(published on-line before print).

Pisano, E. D., C. Kuzmiak, and M. Koomen. 2001. Perspective on digital mammography. *Semin Roentgenol.* 36:195-200.

Salfity, M. F., R. M. Nishikawa, Y. Jiang, and J. Papaioannou. 2003. The use of a priori information in the detection of mammographic microcalcifications to improve their classification. *Med Phys.* 30:823-831.

What is the Required Pixel Size for Digital Mammography?

Robert M. Nishikawa, Yulei Jiang, Ingrid Reiser

Department of Radiology, the University of Chicago

1. Introduction

Pixel size is a critical issue in digital mammography, because it has a direct effect on the cost of systems, image quality, ease of use, and the performance of CAD schemes. It is generally accepted that high spatial resolution is needed in mammography to image accurately the often-subtle signs of cancer and to distinguish cancer from benign lesions. Pixel size has a direct bearing on the limiting resolution of a digital system, since spatial frequencies corresponding to the inverse of the pixel size (or more precisely the sampling frequency) are not imaged correctly. Spatial frequencies higher than the sampling frequency will be aliased. Pixel size also dictates the size of the image, since the image receptor for FFDM must be at least 18x24 cm. At 100-micron pixel size (and sampling distance), the image will be approximately 2048x2560 pixels and this size of image will conveniently fit on a high-resolution display. At 50-micron pixel size, the image will be four times larger. No display monitor exists that can display a full 50-micron FFDM image at full resolution. While high-resolution information exists in such an image, it cannot be easily viewed by a radiologist. Zooming with panning is needed. This could lead to longer reading times and poorer ergonomic designed display systems. Furthermore, the cost of designing and building a 50-micron FFDM system is substantially greater than a 100-micron system, as there are significant technological barriers to using small pixels. Finally, CAD is likely to be an integral part of FFDM systems in the future. The performance of CAD schemes may be influenced by the pixel size of the image.

In this paper, we examine the issue of the required pixel size for digital mammography using results of several observer studies and experiments. Because of the importance of pixel size, there have been a number of studies investigating radiologists' performance and CAD performance on digitized images with different pixel size. These have been either for detecting calcifications or for classifying calcifications. Calcifications are used as the target lesion because it is believed that they provide a better test of the resolution of the system compared to mass lesions. This is true in general, although the margin of masses, in particular spiculations, require high spatial resolution to be imaged accurately.

2. Detection of Calcifications

2.1. Radiologists

There has been one study examining radiologists' ability to detect calcification-like objects in a mammogram. Higashida *et al.* (Higashida *et al.*, 1992) used glass beads 0.125 mm to 0.250 mm in size overlaying 5 cm of breast tissue. These were then imaged either on a screen-film system or on a computer radiography (CR) system, which had a 100-micron pixel size. Nine radiologists read the images. They found that the area under the receiver operating characteristic (ROC) curve was statistically significantly higher for screen-film mammography (0.82 versus 0.72).

In contrast, Nab *et al.* (Nab et al., 1992) compared radiologists ability to detect calcifications in screen-film mammograms and the same mammograms digitized at 100 microns. xx films, xx radiologists. They found not statistically significant difference between digitized film and the original films.

2.2. Computer-Aided Detection (CADe)

Chan *et al.* (Chan et al., 1994) studied the effect of pixel size on the ability of their CADe scheme to detect individual calcifications. They found as the pixel size decreased from 140 to 105 to 70 to 35 microns that the performance of their CADe scheme improved, with a statistically significant improvement between 105 and 35 microns.

3. Classification of Calcifications

3.1. Radiologists

Chan *et al.* (Chan et al., 2001) conducted an observer study to examine the effect of pixel size on radiologists ability to classify mammographic calcifications. They examined pixel sizes of 35, 70, 105, and 140 microns. They found no statistically significant difference in terms of AUC between any pixel sizes.

3.2. Computer-Aided Diagnosis (CADx)

Chan *et al.* (Chan et al., 1996) conducted a study to examine the effect of pixel size on the performance of their computer-aided diagnosis (CADx) scheme. For pixel sizes of 35, 70, 105, and 140 microns, there were no statistically significant differences between any sizes. Further, in their study, they employed a fairly comprehensive set of texture features (using co-occurrence matrices) and morphological features.

This result was also observed by Gavrielides et al., who conducted a smaller study on the comparing the performance of their own CADx scheme as a function of the pixel size of the image. They examined 30, 60, and 90-micron pixels. Considering the results of using manual segmentation and only cases with actual calcifications (as opposed to simulated clusters), their classifier obtained a percent correct of 48, 76 and 80 for 30, 60, and 90 micron pixels. Again, a smaller pixel does not improve performance of CADx schemes.

4. Significance of These Observer Studies

With few exceptions, pixels size of 30 – 100 microns has little effect on performance of radiologists and CADx schemes in classifying calcifications. However, for CADe schemes and perhaps radiologists, smaller pixel size produces better performance. This is contrary to the prevailing thought on pixel size, where it is assumed that to detect a calcification lower spatial resolution can be used because you do not need to know the shape of the calcification in order to detected it, but classification requires higher resolution in order to determine the shape of the calcification.

One possible explanation of this paradox is two fold. First shape is not a reliable diagnostic feature for discriminating between benign and malignant calcification, except on a somewhat gross scale. Second, detection of an object is really a classification task to differentiate an actual calcification from a false positive (caused by noise, artifact, or normal breast structures). We will address this two statements below

5. Importance of Calcification Shape

Jiang et al. (Jiang et al., 1999) performed an observer study to compare radiologists' ability to classify mammographic calcifications with and without a computer aid. In their study, the radiologists read the four standard screening views and magnification views, all using the original films. Magnifying glasses and a bright light were available and the location of the cluster in question was marked on each film. Their CADx scheme used the original screening views only digitized at only 100 microns. That is, radiologists had more information available to them than the CADx had.

The computer used eight features extracted from the mammograms. They were: xxx. Note the only feature related to shape is whether the calcification was linear or not. The results were that computer out performed the radiologist, with AUC of 0.80 compared to xx for the radiologists. This implies that the shape of the calcification is not a necessary diagnostic feature, except on a somewhat gross scale – whether the calcification is linear.

Not only is shape not a necessary feature, but it is an unreliable feature to use to classify calcification because it is difficult to discern the shape from a mammogram. Reiser et al. have shown in a 2AFC experiment that the to discriminate between two different shapes requires the image to have higher signal-to-noise ratio (SNR) than just to detect an object. Note that this experiment does not contract any early experiments since here the independent variable is SNR not pixel size or spatial resolution. They found that the more similar the objects, the higher the SNR need for accurate shape determination. For example, to choose between a star and circle of equal area and contrast, the SNR is 75% higher than the SNR required to detected either object. Their experiment did not include the effect of sampling with finite sized pixels. One would expect that as the size of the object approaches the size of the pixel, an even greater increase in SNR would be need to determine the object's shape. Therefore, for small calcifications that are somewhat subtle, there is unlikely to have sufficient SNR to allow its shape to be determined accurately, unless it is elongated as in a linear or branching calcification. Realistic mammographic backgrounds were not used in the Reiser experiment, so this effect still needs to be studied.

6. Detection is a Classification Task

Detection of a calcification is also a classification task. This is well known by those developing CADe schemes, because the computer has to determine whether a detected object is an actual lesion or a false positive. This is the likely explanation for the result of Chan *et al.* (Chan et al., 1994) Their detection scheme required high spatial frequency information to correctly identify false positives. For example, artifacts on the film, such as emulsion pick off, appear on the mammogram as a small very shape, very high contrast objects -- sharper and higher in contrast than actual calcifications. Therefore, they can easily be identified as an artifact (a potential false positive). However, when the film is digitized, the artifact becomes blurred and its contrast is reduced – the amount is dependent on the pixel size. Therefore, for larger pixel sizes, the artifact can look like a calcification and is more likely to be identified as an actual calcification (i.e., a false positive).

A recent study by xx and Lowe that is consistent with higher spatial resolution for determining false positives. In their study, they examined eye-tracking data of radiologists reading mammograms. They then analyzed essentially the spatial frequency content of areas that radiologists dwelled upon. They found that for cases with no actual lesions, the radiologist examined areas with more high spatial frequency information, than in cases were a lesion was present.

7. Pixel Size versus SNR

In a digital image, the ultimate limit to detecting or characterizing an object is SNR not spatial resolution, since, in a digital image, image processing can be used to increase high spatial frequency information. This assumes that the object is bigger than the pixel size. Therefore, it is more important to examine SNR as a function of spatial frequency than to examine pixel size exclusively. This can be done using the concept of noise equivalent quanta (NEQ) or to determine a task-based SNR (ICRU, 1996). We hypothesize that the mid-frequency information (between 1 cycle/mm and 3-4 cycles/mm) is more important factor than pixel size by itself.

8. Conclusions

Contrary to conventional wisdom, accurate classification of microcalcifications can be performed using 100 micron pixel size, whereas for computer-aided detection of microcalcifications smaller pixel sizes will improve performance by reducing the false positive rate. The larger pixel size for classification implies that the shape of microcalcifications is not a reliable feature for differentiating benign and malignant calcifications. Finally, we believe that high SNR at intermediate spatial frequencies is more important than pixel size in determining performance of radiologists and computers in detecting and classifying breast lesions, but this still needs to be demonstrated.

8. References

Chan, H.-P., L. T. Niklason, D. M. Ikeda, K. L. Lam, & D. D. Adler. 1994. Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications. *Medical Physics*, 21:1203-1211.

Chan, H.-P., B. Sahiner, N. Petrick, K. L. Lam, & M. A. Helvie. 1996. Effects of pixel size on classification of microcalcifications on digitized mammograms. *Proc. SPIE*, 2710:30-41.

Chan, H. P., M. A. Helvie, N. Petrick, B. Sahiner, D. D. Adler, C. Paramagul, M. A. Roubidoux, C. E. Blane, L. K. Joynt, T. E. Wilson, L. M. Hadjiiski, & M. M. Goodsitt. 2001. Digital mammography: observer performance study of the effects of pixel size on the characterization of malignant and benign microcalcifications. *Academic Radiology*, 8:454-66.

Higashida, Y., N. Moribe, K. Morita, N. Katsuda, M. Hatemura, T. Takada, M. Takahashi, & J. Yamashita. 1992. Detection of subtle microcalcifications: comparison of computed radiography and screen-film mammography. *Radiology*, 183:483-486.

ICRU. 1996. *Medical Imaging - The Assessment of Image Quality. ICRU Report 54.* Bethesda, Maryland: International Commission on Radiation Units and Measurements.

Jiang, Y., R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, & K. Doi. 1999. Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Radiology*, 6:22-33.

Nab, H. W., N. Karssemeijer, L. Vanerning, & J. Hendriks. 1992. Comparison of Digital and Conventional Mammography - a ROC Study of 270 Mammograms. *Medical Informatics*, 17:125-131.

International Congress Series 1268 (2004) 886–889

www.ics-elsevier.com

Comparison of two methods of adding jitter to artificial neural network training

R.M. Zur*, Y. Jiang, C.E. Metz

Department of Radiology MC2026, The University of Chicago, 5841 South Maryland Avenue, 60637, Chicago, IL, USA

Abstract. We compare two methods of training artificial neural networks (ANNs) that potentially reduce the risk of the neural network overfitting the training data set. We refer to these methods as training with jitter. In one method of training with jitter, a new random noise vector is added to each training-data vector between successive iterations. In this work, we propose a different method of training with jitter, in which instead of adding different random noise vectors between iterations, a number of random vectors are used to expand the training data set prior to training. This artificially expanded data set is then used to train the artificial neural network in the conventional manner. These two methods are compared to the conventional method of training artificial neural networks. We find that although training with a single expanded training data set does increase the performance of the neural networks, overfitting can still occur after a large number of training iterations. © 2004 CARS and Elsevier B.V. All rights reserved.

Keywords: Artificial neural networks; Computer-aided diagnosis; Classification

1. Introduction

Artificial neural networks (ANNs) are popular classification algorithms in computer-aided diagnosis because of their ability to "learn" classification rules from a set of training data. However, ANNs run the risk of learning the training data too closely, resulting in networks that do not perform well when presented with new, unknown data. This problem is known as overfitting, and numerous methods exist to guard against it [1]. For example, limiting the number of hidden layers and hidden nodes reduces the chances of overfitting.

Training with jitter may help an ANN generalize well on new data [2,3]. Because each training data set is a finite sample from the underlying population, it may not represent that population accurately. Training with jitter increases the size of the training data set by supplementing it with additional artificial data that is similar to, but different from, the real

* Corresponding author. Tel.: +1-773-834-5094; fax: +1-773-702-0371.
E-mail address: zur@uchicago.edu (R.M. Zur).

0531-5131/ © 2004 CARS and Elsevier B.V. All rights reserved.
doi:10.1016/j.ics.2004.03.238

training data. This can cause the data set to appear smoother to the ANN. If this smooth, though partially artificial, data set approximates the underlying population, then the ANN may perform better when presented with new data.

Previously, we trained ANNs with jitter by adding a random noise vector to each training input vector. Before each training iteration, a new, zero-mean random noise vector was drawn from a Gaussian distribution. In this way, after a large number of iterations, the artificial neural network was presented with a large number of "jittered" training cases, in effect increasing the size of the training data set. In the work repeated here, we compare this method with a simpler version. Rather than select a new random noise vector for each iteration, we instead add a large number of artificial cases before training the neural network in the conventional manner.

2. Method

We trained eight artificial neural networks, each with different initial weights, on simulated data. Each neural network had two input nodes, a single hidden layer with 10 hidden nodes, and a single output node. Although 10 hidden nodes may have been too many for our simple problem, we were attempting to investigate overfitting, and more free parameters in the algorithm makes overfitting more likely. Training was stopped after 12,500 iterations, because by then overfitting was usually evident. The performance of each ANN was tested using a large test data set drawn from the same population. The results were characterized using receiver operating characteristic (ROC) analysis. The performance was indexed by A_Z, the area under the maximum-likelihood-fit binormal ROC curve.

Our first method of training with jitter required the addition of a random noise vector to each training input vector. After each iteration, a new random noise vector was added to the original input vector. Thus, as the artificial neural network is being trained the training input vector moves, or jitters, around the original input vector in feature space. The random noise vectors were drawn from a zero-mean Gaussian distribution, the variance of which was a parameter specified by the user. We investigated various variance values to evaluate the dependence of the performance on this parameter.

To use our new method to train artificial neural networks, we added many independent random noise vectors into the training data set prior to training. Thus, a single large, partially artificial data set was used to train the ANNs in the conventional way.

Simulated data sets were used to train and test our artificial neural networks. We used XOR distributions because they are a simple-to-understand, two-dimensional distribution that has a nonlinear optimal decision boundary. The normal and abnormal distributions were each composed of two Gaussian probability densities. The means of the four Gaussians were arranged in a square, with the normal cases drawn from Gaussian distributions centered at the bottom-left and top-right corners while the abnormal cases drawn from Gaussian distributions centered at the bottom-right and top-left corners. Because the XOR distributions are two-dimensional, the data was comprised of two features.

For training, 10 normal and 10 abnormal cases were drawn from the underlying distributions. Although this is a small data set, we expect a small number of training cases

Table 1
Results of training ANNs in the conventional manner, with continuously changing jitter, and with our new method of adding jitter prior to conventional training

Training method	Number of training cases (training cases + "jitter" cases)	Average maximum performance [A_Z]	Jitter variance
Conventional	20	0.723	N/A
Jitter	20	0.770	0.08
Prior-to-training	20 + 20	0.758	0.04
Jitter	20 + 60	0.753	0.09
	20 + 140	0.750	0.11
	20 + 300	0.757	0.06
	20 + 620	0.754	0.10
	20 + 1260	0.747	0.07

to inadequately represent the underlying population, thus allowing the ANNs to overfit more easily. For testing, 1000 normal and 1000 abnormal cases were drawn from the same pair of bimodal underlying distributions. The large number of testing cases was employed to ensure that a precise estimate of performance was obtained.

3. Results

For eight ANNs trained in the conventional manner, without jitter, the average maximum A_Z was 0.723 on the testing data set. By training with the continually added jitter, the average maximum performance of the ANNs increased to an A_Z of 0.770 when jitter of variance 0.08 was added. Because the A_Z value fluctuated with successive training iterations, we fitted a 9th-order polynomial to smooth the results. The maximum of the polynomial was taken as the maximum performance.

The results are summarized in Table 1. The same 20 training cases were used for training in the conventional method as well as the original method of training with jitter. For adding jitter using our new method, we used the original 20 training cases with 20, 60, 140, etc. artificial "jitter" cases added prior to training. For the results from adding jitter, we show the maximum average A_Z attained as well as the jitter variance that attained that performance.

4. Conclusions

In conclusion, we find that adding jitter as a preliminary step to neural network training can increase the performance of ANNs. Training with continually changing jitter seems to give better performance, but both show improvement over training without jitter. In fact, in our simple simulation, even a small increase in the number of training samples prior to training the networks provided an increase in performance. The method of training with jitter added prior to training, however, showed evidence of overfitting after a large number of training iterations.

While this result may be valid in general, our work is still preliminary. Further investigation of more realistic conditions, such as larger data sets with more features and more complicated distributions, is required.

References

[1] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford Univ. Press, New York, 1995.

[2] L. Holmström, P. Koistinen, Using additive noise in back-propagation training, IEEE Trans. Neural Netw. 3 (1) (1992) 24–38.

[3] G. An, The effects of adding noise during backpropagation training on a generalization performance, Neural Comput. 8 (1996) 643–674.